AD NUMBER

**AD477349**

LIMITATION CHANGES

TO:

**Approved for public release; distribution is unlimited.**

FROM:

**Distribution authorized to U.S. Gov't. agencies and their contractors;
Administrative/Operational Use; NOV 1965. Other requests shall be referred to Electronic Systems Div., Hanscom AFB, MA.**

AUTHORITY

**ESD ltr 30 Sep 1968**

ESD-TR-65-577

(Final Report)

# APPLICATION OF QUEUING THEORY TO INFORMATION SYSTEM DESIGN

Dovid Nee

November 1965



DIRECTORATE OF COMPUTERS
ELECTRONIC SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE
L. G. Hanscom Field, Bedford, Massochusetts




(Prepared under Contract No. AF 19(628)-4341 by the Stanford Reseorch Institute, Menlo Park, Colifornio)

## LEGAL NOTICE

When U. S. Government drawings, specifications or other data are used for any purpose other than a definitely related government procurement operation, the government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

## OTHER NOTICES

Do not return this copy. Retain or destroy.

**The following notice applies to any unclassified (including originally classified and now declassified) technical reports released to "qualified U.S. contractors" under the provisions of DoD Directive 5230.25, Withholding of Unclassified Technical Data From Public Disclosure.**

NOTICE TO ACCOMPANY THE DISSEMINATION OF EXPORT-CONTROLLED TECHNICAL DATA

1. Export of information contained herein, which includes, in some circumstances, release to foreign nationals within the United States, without first obtaining approval or license from the Department of State for items controlled by the International Traffic in Arms Regulations (ITAR), or the Department of Commerce for items controlled by the Export Administration Regulations (EAR), may constitute a violation of law.

2. Under 22 U.S.C. 2778 the penalty for unlawful export of items or information controlled under the ITAR is up to ten years imprisonment, or a fine of $1,000,000, or both. Under 50 U.S.C., Appendix 2410, the penalty for unlawful export of items or information controlled under the EAR is a fine of up to $1,000,000, or five times the value of the exports, whichever is greater; or for an individual, imprisonment of up to 10 years, or a fine of up to $250,000, or both.

3. In accordance with your certification that establishes you as a "qualified U.S. Contractor", unauthorized dissemination of this information is prohibited and may result in disqualification as a qualified U.S. contractor, and may be considered in determining your eligibility for future contracts with the Department of Defense.

4. The U.S. Government assumes no liability for direct patent infringement, or contributory patent infringement or misuse of technical data.

5. The U.S. Government does not warrant the adequacy, accuracy, currency, or completeness of the technical data.

6. The U.S. Government assumes no liability for loss, damage, or injury resulting from manufacture or use for any purpose of any product, article, system, or material involving reliance upon any or all technical data furnished in response to the request for technical data.

7. If the technical data furnished by the Government will be used for commercial manufacturing or other profit potential, a license for such use may be necessary. Any payments made in support of the request for data do not include or involve any license rights.

8. A copy of this notice shall be provided with any partial or complete reproduction of these data that are provided to qualified U.S. contractors.

DESTRUCTION NOTICE

For classified documents, follow the procedure in DoD 5220.22-M, National Industrial Security Program, Operating Manual, Chapter 5, Section 7, or DoD 5200.1-R, Information Security Program Regulation, Chapter 6, Section 7. For unclassified, limited documents, destroy by any method that will prevent disclosure of contents or reconstruction of the document.

ESD-TR-65-577

(Final Report)

# APPLICATION OF QUEUING THEORY TO INFORMATION SYSTEM DESIGN

David Nee

November 1965

DIRECTORATE OF COMPUTERS
ELECTRONIC SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE
L. G. Hanscam Field, Bedford, Massachusetts




(Prepared under Contract No. AF 19(628)-4341 by the Stanford Research Institute, Menlo Park, California)

## FOREWORD

This final report on application of Queuing Theory Information Systems Design was prepared by Stanford Research Institute, Menlo Park, California, on Air Force Contract AF 19(628)-4341 as SRI Project 5187. The principal investigator is D. Nee.

The author wishes to acknowledge the contribution to this project by K. Quisel on the study of simulation strategies and A. Beja on the investigation of the variance reduction methods.

## REVIEW AND APPROVAL

This technical report has been reviewed and is approved.

CHARLES A. LAUSTRUP, Lt. Col., USAF
Directorate of Computers
Deputy for Engineering and Technology

## ABSTRACT

This research project was undertaken to evaluate the applicability of queuing theory—in particular the priority queuing theory—to the evaluation of priority queuing situations in military information systems.

The queuing theory has been applied to the evaluation of the queuing situations in the 473-L System. The waiting-time distribution for a single-server, head-of-the-line, priority queuing model has been evaluated. The application of "Variance Reduction Method" to the Simulation of Priority Queuing Systems has been investigated. Finally, guides and procedures for the application of queuing models in structuring information systems were outlined.

# CONTENTS

# ILLUSTRATIONS

# TABLES

# I INTRODUCTION

## A. OBJECTIVES

The objectives of the research effort described in this report are:

(1) To evaluate the applicability of queuing theory to evaluation of priority queuing situations in military information systems.

(2) Develop additions and/or modifications to the existing priority queuing techniques—those that are deemed necessary from the application studies—in sufficient detail for use as a system design aid.

(3) To establish guides, procedures, and principles for the application of appropriate queuing models in structuring information systems.

## B. SCOPE

The 473-L Simulation Model has been chosen as the sample situation for use in the evaluation of the applicability of queuing theory to the evaluation of priority queuing situations in military information systems. The analytical analysis of the 473-L Model is presented in Appendix A. Results of the analytical investigation have been compared with that obtained from simulation and are presented in Sec. II.

Usually the analysis of priority queuing is limited to the mean of the operational measures. An analytical expression for the priority queuing waiting-time distribution has been developed during the previous research effort. It is shown in Sec. III that this waiting-time distribution can be approximated by a simpler expression when the system loading and the relative loading of the higher priority to the lower priority customers meet certain requirements.

Although it is possible to analytically investigate the queuing situations in information systems, this usually involves extensive idealizations of the system parameters. The results obtained are usually limited to the mean of the measures. This has been shown in Sec. II. For more complex information systems, where

(1) the inter-arrival interval distributions or the service time distributions are not exponential, and

(2) where queuing discipline is of the priority type and the service center contains multiple parallel servers, the only practical technique for system evaluation is the simulation modeling technique.

It is therefore highly desirable to develop simulation strategies that can be used to increase the effectiveness of this technique. Three aspects of the simulation strategies have been investigated in Sec. IV. These are the stopping rules and the chopping rules of the simulation experiment, and the application of "Variance Reduction Methods" to the Simulation of Priority Queuing Systems.

Finally, the guides and procedures for the application of the Queuing Models are outlined in Sec. V.

# II RESULTS OF ANALYTICAL ANALYSIS OF 473-L SYSTEM

## A. INTRODUCTION

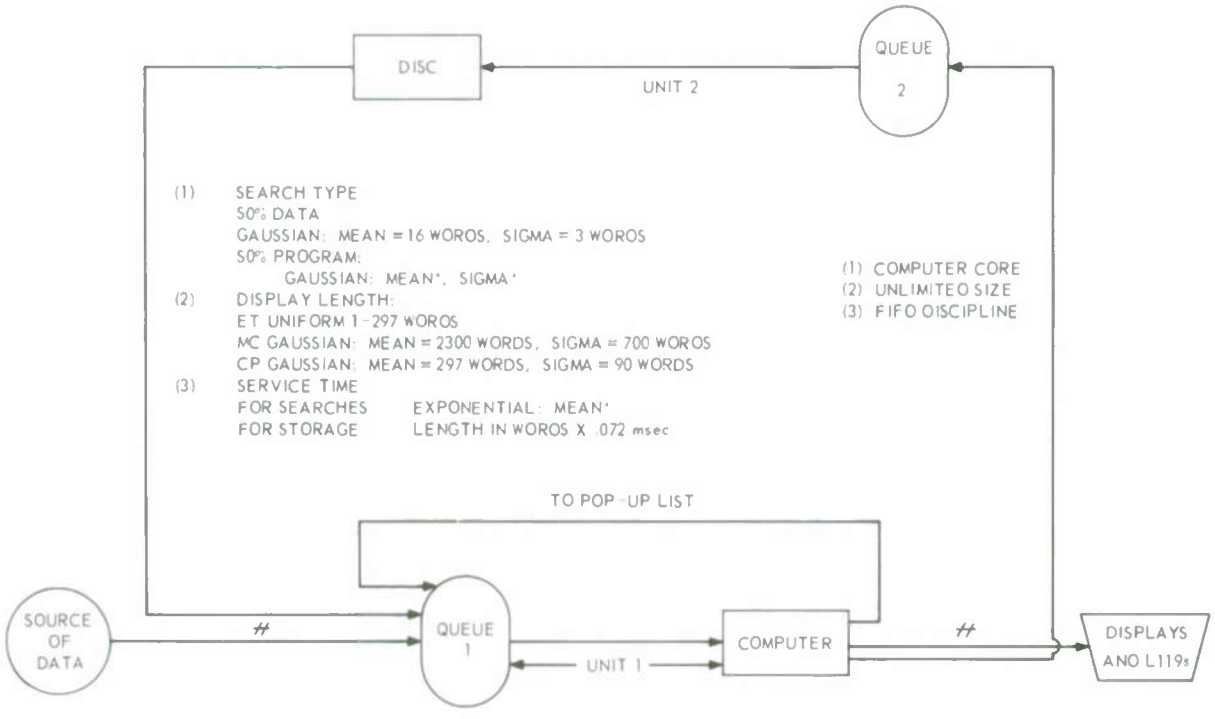
The principal purpose of the queuing analysis of the 473-L system is to investigate the possibility of applying the queuing theory to obtain some of the queuing measures of a real information system. The 473-L system model has been chosen for this purpose because a simulation model of this system has been constructed and simulated results of this system exist.[1*] Thus, it is possible to check the analytical results with the simulated results.

## B. DESCRIPTION OF 473-L SYSTEM

In this section a brief description of the 473-L Queuing model is given. For a more detailed description see Appendix A or Reference 1. Figure II-1 shows the 473-L Queuing Simulation Model. Requests for service arrive at Buffer 1, $Q$-1, from the source of data. The requests are categorized into three priority classes. The queuing discipline at Queue 1 is that of interrupt priority type. All requests require an initial service by the computer. Some of the requests require further service(s) from the computer. Each further service by the computer is preceeded by a disc service which is to retrieve the desired information. Those requests that require further service(s) from the computer would have to enter the Buffer 2, $Q$-2, to wait for the retrieval operation. The queuing discipline at $Q$-2 is of the ordered queue type. When the required number of computer services have been completed, an output message is sent to the display.

---

* References are listed at the end of the report.




SOURCE:

(1) TIME BETWEEN ARRIVALS
EXPONENTIAL: MEAN*
(2) THREE LEVEL PRIORITY
63.2% LOW
26.3% MEDIUM
10.5% HIGH
(3) DISPLAY OR MESSAGE TYPE
80% ET OR L119
10% MC
10% CP
(4) REQUEST LENGTH
50% 2 WOROS
50% UNIFORM 1-297 WOROS

BUFFER:

(1) COMPUTER CORE
(2) UNLIMITED SIZE
(3) POP-UP LIST LENGTH
(a) FOR TASK IN PROCESS
# SEARCHES X 5 WORDS
(b) FOR REQUESTS IN QUEUE
# REQUESTS X 1 WORD
(4) PRIORITY ORDERING IN BUFFER

COMPUTER:

(1) SERVICE TIME
EXPONENTIAL: MEAN*
(2) SEARCHES
25% 0
75% GAUSSIAN: MEAN*
SIGMA*
(3) ALL BUFFER INPUTS CAUSE HOUSEKEEPING INTERRUPT TIME 1 msec

OUTPUT:

READOUT RATE
ET = 0
MC = 5.6 msec X # WORDS
CP = 560 msec X # WORDS

# RESPONSE TIME IS TIME DELAY FROM POINT A TO POINT B

* PARAMETERS VARIED
(1) MEAN TIME BETWEEN ARRIVALS: 2, 3, 4, 5, 6, 7, 9, 12 SEC
(2) COMPUTER SERVICE TIME: 40, 50, 65, 75, 80, 90, 180, 200, 220 msec
(3) SEARCHES: MEAN = 9, 15, SIGMA = 3, 5
(4) PROGRAM LENGTH: MEAN = 1000, 2500 WOROS, SIGMA = 300, 600
(5) DISC SERVICE TIME: 90, 180, 360, 540 msec

TB-5187-2

FIG. II-1 473-L SIMULATION MODEL

## C. RESULTS OF THE ANALYSIS

The analytical expressions for some of the queuing measures are

(1) the utility of the computer $\rho_c$

(2) the utility of the disc, $\rho_d$

(3) the mean response time of all requests, $T_r$

(4) the mean response time of each priority class of requests, $T_i$, and

(5) the mean storage size for $Q$-1 and $Q$-2, $L_{Q1}$ and $L_{Q2}$.

The utility of the computer is defined as the percentage of time the computer spent in servicing requests. Similarly, the utility of the disc is the percentage of time that the disc spent in servicing requests. The mean response time is defined as the mean elapsed time between arrival of the new request at $Q$-1 to the time when it leaves the computer after final service by the computer. The mean storage size is defined as the mean number of words in the buffers.

Detailed derivations of these expressions are contained in Appendix A. The expression for each of the above queuing measures are listed below:

(1) The utility of the computer

$$\rho_c = \frac{\lambda h_c [1 + (1 - X)(n + 1)]}{1 - \lambda\{h_c(1 - X)(1 - Z) + h_h[1 + (1 - X)(n + 1)Z]\}} \tag{1}$$

(2) The utility of the disc

$$\rho_d = \lambda(1 - X)\{[n + 1 + \rho_c(1 - Z)]h_d + \rho_c(1 - Z)h_s\} \tag{2}$$

(3) The mean response time (sec) of all requests

$$T_r = \frac{Xh_c}{1 - \rho_c} + (1 - X)\left\{(n + 1)\left[\frac{h_c}{1 - \rho_c} + \frac{h_d}{1 - \rho_d}\right] + \rho_c(1 - Z)\left[\frac{h_c}{1 - \rho_c} + \frac{h_d + h_s}{1 - \rho_d}\right]\right\} \tag{3}$$

(4) The mean response time (sec) of the $i$th priority request

$$T_i = T'_i + (1 - X)[(n + 1)(T'_{i+3} + T_d) + \rho_c \sum_{j=1}^{i} Y_i (T'_{i+3} + T_p)] \qquad (4)$$

where

$$T'_i = h_c / [(1 - P_{i-1})(1 - P_i)]$$

$$P_i = \sum_{k=1}^{i} \rho_i$$

$$\rho_i = \lambda h_c Y_i \quad , \quad i = 1, 2, 3$$

$$\rho_i = \lambda h_c Y_i (1 - X)\left(n + 1 + \rho_c \sum_{j=1}^{i-3} Y_i\right) \quad , \quad i = 4, 5, 6 \; .$$

$$T_d = h_{se} + (h_d \rho_d)/(1 - \rho_d)$$

$$T_p = h_{sp} + (h_d \rho_d)/(1 - \rho_d)$$

(5) The mean storage size (words) has been developed under two sets of storage rule, Rule 1 and Rule 2. See Appendix A for more detailed explanation of these rules.

The mean storage size for $Q$-1 and $Q$-2, according to Rule 1, are given as

$$L_{Q1} = \frac{M_n(1 - X)\left\{n\left[\frac{n + 1}{2} M_p + V_n M_{ds} + (1 - V_n) M_{ps}\right] + M_c + [M_n + (n + 2) M_p]\rho_c (1 - Z)\right\}}{(1 - \rho_c)\{1 + (1 - X)[n + 1 + 2\rho_c (1 - Z)]\}} \qquad (5)$$

and

$$L_{Q2} = \frac{\frac{(n + 1)(n + 2)}{2} M_p + n[V_n M_{ds} + (1 - V_n) M_{ps}] + [M_n + (n + 3) M_p]\rho_c (1 - Z)}{(1 - \rho_d)[n + 1 + 2\rho_c (1 - Z)]} \; . \qquad (6)$$

The mean storage size for $Q$-1 and $Q$-2, according to Rule 2, are given as

$$L_{Q1} = \frac{M_n + (1 - X)\left\{n\left[\frac{n + 1}{2} M_p + V_n M_{ds} + (1 - V_n) M_{ps}\right] + M_c + (n + 2)(M_n + M_p)\rho_c (1 - Z)\right\}}{(1 - \rho_c)\{1 + (1 - X)[n + 1 + 2\rho_c (1 - Z)]\}} \; , \qquad (7)$$

and

$$L_{Q2} = \frac{(n+1)(n+2)M_p + n[V_n M_{ds} + (1-V_n)M_{ps}] + (M_n + M_p)(n+3)\rho_c(1-Z)}{(1-\rho_d)[n+1+2\rho_c(1-Z)]} \quad (8)$$

where

$\lambda$ = The mean arrival rate of requests from the source of data.

$h_c$ = The mean service time of the computer.

$X$ = The percentage of new requests that require no disc search.

$h_n$ = The housekeeping time due to an interrupt.

$Z$ = The probability that a request, upon its arrival at $Q$-1, finds that the request that is being serviced by the computer has a lower or equal priority.

$h_s$ = The mean service time for the disc search request.

$h_{se}$ = The mean service time for the data disc search request.

$h_{sp}$ = The mean service time for the program disc search request.

$n$ = The mean number of disc searches required by a new request.

$Y_i$ = The percentage of new requests that are $i$th priority requests.

$M_n$ = The mean length of a new request.

$M_p$ = The mean length of a disc search address.

$M_{ds}$ = The mean length of the message retrieved by a data disc search request.

$M_{ps}$ = The mean length of the message retrieved by a program disc search request.

$V_n$ = The proportion of $n$ disc searches that are the data disc searches.

$M_c$ = The mean length of a display message.

The results of the simulation for twenty-four cases are listed in Table I, as extracted from Reference 1.

Table I

COMPUTED AND SIMULATED RESULTS OF THE 473L SIMULATION MODEL

| Case No. | Mean Arrival--$\lambda$ (requests/sec) | Mean Computer Service Time, $h_c$-msec | Mean Disc Service Time, $h_d$-msec | Mean Disc Search, N | | Computer Utilization, % | | Disc Utilization, % | | Mean Response Time, $T_r$ (sec) | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Specified | Simulated | Computed | Simulated | Computed | Simulated | Computed | Simulated |
| 1 | 1/3 | 50 | 180 | 9 | 8.45 | 12.35 | 12.7 | 42.97 | 40.45 | 2.58 | 2.46 |
| 2 | 1/3 | 50 | 180 | 9 | 8.45 | 12.35 | 12.7 | 42.97 | 40.45 | 2.58 | 2.46 |
| 3 | 1/3 | 50 | 180 | 15 | 14.9 | 20.51 | 22.9 | 72.36 | 77.24 | 10.34 | 10.45 |
| 4 | 1/3 | 50 | 180 | 15 | 14.9 | 20.51 | 22.9 | 72.36 | 77.24 | 10.34 | 10.45 |
| 5 | 1/3 | 50 | 90 | 15 | 14.7 | 20.26 | 22.37 | 35.73 | 36.46 | 2.47 | 2.50 |
| 6 | 1/6 | 50 | 360 | 15 | 14.67 | 10.06 | 10.84 | 70.89 | 71.77 | 15.23 | 13.42 |
| 7 | 1/7 | 50 | 360 | 15 | 14.45 | 8.49 | 9.09 | 59.87 | 60.07 | 11.15 | 11.82 |
| 8 | 1/12 | 50 | 360 | 15 | 14.56 | 4.98 | 5.28 | 35.10 | 35.09 | 7.12 | 7.49 |
| 9 | 1/9 | 50 | 540 | 15 | 14.53 | 6.63 | 7.29 | 70.14 | 71.85 | 23.07 | 21.97 |
| 10 | 1/10 | 50 | 540 | 15 | 14.55 | 5.98 | 6.00 | 63.16 | 59.21 | 16.12 | 16.97 |
| 11 | 1/3 | 65 | 180 | 15 | 14.29 | 25.73 | 26.41 | 69.74 | 67.70 | 7.52 | 6.76 |
| 12 | 1/3 | 75 | 180 | 9 | 8.77 | 19.23 | 20.97 | 44.70 | 43.89 | 3.16 | 3.33 |
| 13 | 1/3 | 75 | 180 | 15 | 14.87 | 30.85 | 32.83 | 72.57 | 73.67 | 9.65 | 9.66 |
| 14 | 1/3 | 200 | 180 | 9 | 8.45 | 50.84 | 60.31 | 44.65 | 48.46 | 6.47 | 6.40 |
| 15 | 1/3 | 90 | 180 | 15 | 14.33 | 35.89 | 40.21 | 70.40 | 73.67 | 9.83 | 9.03 |
| 16 | 1/3 | 180 | 180 | 9 | 8.52 | 45.91 | 46.59 | 44.48 | 41.18 | 4.63 | 4.87 |
| 17 | 1/3 | 220 | 180 | 9 | 8.48 | 56.32 | 59.51 | 44.76 | 44.47 | 6.60 | 6.60 |
| 18 | 1/3 | 50 | 360 | 15 | 14.70 | 20.26 | 15.31 | 142.36 | 99.93 | 6101.86 | 663.00 |
| 19 | 1/4.5 | 50 | 360 | 15 | 14.51 | 13.30 | 14.16 | 93.72 | 92.60 | 57.73 | 48.50 |
| 20 | 1/9 | 50 | 360 | 15 | 14.38 | 6.57 | 7.17 | 46.31 | 47.58 | 8.58 | 8.67 |
| 21 | 1/12 | 50 | 540 | 15 | 14.54 | 4.97 | 5.23 | 50.55 | 51.58 | 13.66 | 14.67 |
| 22 | 1/2 | 50 | 90 | 15 | 14.51 | 30.18 | 31.70 | 53.21 | 51.90 | 3.10 | 3.20 |
| 23 | 1/3 | 80 | 180 | 15 | 14.33 | 31.84 | 34.10 | 70.19 | 69.40 | 8.33 | 8.00 |
| 24 | 1/3 | 40 | 180 | 15 | 14.16 | 15.63 | 16.60 | 68.80 | 68.60 | 7.14 | 7.57 |

The queuing measures, computer utility $\rho_c$, the disc utility $\rho_d$, and the response time $T_r$, are computed, using Eqs. (1), (2), and (3), respectively, for the twenty-four cases. The computed values are listed in Table II together with the simulated values. The $\rho_c$ is computed on the basis of simulated mean number of disc search. The $\rho_d$ is computed on the basis of simulated computer utilization and simulated mean number of disc search. Finally, the simulated $n$, $\rho_c$, and $\rho_d$ were used to compute the $T_r$. The reason for using the simulated $n$, $\rho_c$, and $\rho_d$ in computing each of the queuing measures is to make the system environments of the analytical model match as closely as possible that of the simulation model.

Table III is a listing of the simulated and computed mean response time for the $i$th priority request, $T_i$, for the above cited twenty-four cases. Equation (4) was used to compute the $T_i$. Once again, the simulated $n$, $\rho_c$ and $\rho_d$ were used in the computation.

Equations (5) and (6) were used to compute the mean storage size in $Q$-1 and $Q$-2 for the twenty-four cases. These are listed in the second and third column of Table IV. The fourth column is the sum total of the first two columns. The $L_{Q1}$ and $L_{Q2}$ listed in the fifth and sixth columns were computed according to Eqs. (7) and (8). Their sums are listed in the seventh column. The simulated results are listed in the eighth column. The percent deviation of the simulated results from the computed results, according to Rule 1 and Rule 2, are tabulated in the ninth and tenth columns.

## D. DISCUSSION

It is noted in Table II that the analytical results are generally in agreement with the simulated results, except for a few specific cases. The most noticeable case is Case 18, the only one where the computed $\rho_c$ is higher than the simulated $\rho_c$. Upon further examination, it is seen Case 18 is the only one where the disc is saturated. The computed disc utility is 142.36%, which should be interpreted as the offered load at $Q$-2, but the actual utility of the disc is 100%. Since the disc is saturated, the mean rate at which the request returns to the $Q$-1 after the disc search becomes independent of the arrival rate of new requests. This rate is the mean disc search rate, which is approximately $1/h_d$. Thus, the mean arrival rate

Table 11

COMPUTED AND SIMULATED RESULTS OF THE 473-L SIMULATION MODEL

| CASE NO. | MEAN ARRIVAL—$\lambda$ (requests/sec) | MEAN COMPUTER SERVICE TIME, $h_c$-msec | MEAN DISC SERVICE TIME, $h_d$-msec | MEAN DISC SEARCH, $N$ | | COMPUTER UTILIZATION, % | | DISC UTILIZATION, % | | MEAN RESPONSE TIME, $T_r$ (sec) | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Specified | Simulated | Computed | Simulated | Computed | Simulated | Computed | Simulated |
| 1 | 1/3 | 50 | 180 | 9 | 8.45 | 12.35 | 12.7 | 42.97 | 40.45 | 2.58 | 2.46 |
| 2 | 1/3 | 50 | 180 | 9 | 8.45 | 12.35 | 12.7 | 42.97 | 40.45 | 2.58 | 2.46 |
| 3 | 1/3 | 50 | 180 | 15 | 14.9 | 20.51 | 22.9 | 72.36 | 77.24 | 10.34 | 10.45 |
| 4 | 1/3 | 50 | 180 | 15 | 14.9 | 20.51 | 22.9 | 72.36 | 77.24 | 10.34 | 10.45 |
| 5 | 1/3 | 50 | 90 | 15 | 14.7 | 20.26 | 22.37 | 35.73 | 36.46 | 2.47 | 2.50 |
| 6 | 1/6 | 50 | 360 | 15 | 14.67 | 10.06 | 10.84 | 70.89 | 71.77 | 15.23 | 13.42 |
| 7 | 1/7 | 50 | 360 | 15 | 14.45 | 8.49 | 9.09 | 59.87 | 60.07 | 11.15 | 11.82 |
| 8 | 1/12 | 50 | 360 | 15 | 14.56 | 4.98 | 5.28 | 35.10 | 35.09 | 7.12 | 7.49 |
| 9 | 1/9 | 50 | 540 | 15 | 14.53 | 6.63 | 7.29 | 70.14 | 71.85 | 23.07 | 21.97 |
| 10 | 1/10 | 50 | 540 | 15 | 14.55 | 5.98 | 6.00 | 63.16 | 59.21 | 16.12 | 16.97 |
| 11 | 1/3 | 65 | 180 | 15 | 14.29 | 25.73 | 26.41 | 69.74 | 67.70 | 7.52 | 6.76 |
| 12 | 1/3 | 75 | 180 | 9 | 8.77 | 19.23 | 20.97 | 44.70 | 43.89 | 3.16 | 3.33 |
| 13 | 1/3 | 75 | 180 | 15 | 14.87 | 30.85 | 32.83 | 72.57 | 73.67 | 9.65 | 9.66 |
| 14 | 1/3 | 200 | 180 | 9 | 8.45 | 50.84 | 60.31 | 44.65 | 48.46 | 6.47 | 6.40 |
| 15 | 1/3 | 90 | 180 | 15 | 14.33 | 35.89 | 40.21 | 70.40 | 73.67 | 9.83 | 9.03 |
| 16 | 1/3 | 180 | 180 | 9 | 8.52 | 45.91 | 46.59 | 44.48 | 41.18 | 4.63 | 4.87 |
| 17 | 1/3 | 220 | 180 | 9 | 8.48 | 56.32 | 59.51 | 44.76 | 44.47 | 6.60 | 6.60 |
| 18 | 1/3 | 50 | 360 | 15 | 14.70 | 20.26 | 15.31 | 142.36 | 99.93 | 6101.86 | 663.00 |
| 19 | 1/4.5 | 50 | 360 | 15 | 14.51 | 13.30 | 14.16 | 93.72 | 92.60 | 57.73 | 48.50 |
| 20 | 1/9 | 50 | 360 | 15 | 14.38 | 6.57 | 7.17 | 46.31 | 47.58 | 8.58 | 8.67 |
| 21 | 1/12 | 50 | 540 | 15 | 14.54 | 4.97 | 5.23 | 50.55 | 51.58 | 13.66 | 14.67 |
| 22 | 1/2 | 50 | 90 | 15 | 14.51 | 30.18 | 31.70 | 53.21 | 51.90 | 3.10 | 3.20 |
| 23 | 1/3 | 80 | 180 | 15 | 14.33 | 31.84 | 34.10 | 70.19 | 69.40 | 8.33 | 8.00 |
| 24 | 1/3 | 40 | 180 | 15 | 14.16 | 15.63 | 16.60 | 68.80 | 68.60 | 7.14 | 7.57 |

Table III

COMPUTED AND SIMULATED MEAN RESPONSE TIME

| CASE NO. | MEAN RESPONSE TIME (sec) PRIORITY | | | | | |
|---|---|---|---|---|---|---|
| | 1st | | 2nd | | 3rd | |
| | Computed | Simulated | Computed | Simulated | Computed | Simulated |
| 1 | 2.56 | 2.25 | 2.58 | 2.39 | 2.64 | 2.54 |
| 2 | 2.56 | 2.25 | 2.58 | 2.39 | 2.64 | 2.54 |
| 3 | 10.11 | 11.29 | 10.21 | 10.33 | 10.43 | 10.38 |
| 4 | 10.11 | 11.29 | 10.21 | 10.33 | 10.43 | 10.38 |
| 5 | 2.34 | 2.16 | 2.40 | 2.25 | 2.56 | 2.68 |
| 6 | 15.15 | 12.26 | 15.18 | 12.28 | 15.29 | 13.82 |
| 7 | 11.10 | 12.51 | 11.13 | 11.59 | 11.34 | 11.82 |
| 8 | 7.06 | 6.10 | 7.03 | 6.93 | 7.11 | 7.94 |
| 9 | 22.96 | 20.79 | 23.03 | 21.53 | 23.13 | 22.36 |
| 10 | 16.10 | 15.59 | 16.11 | 15.78 | 16.16 | 17.81 |
| 11 | 7.27 | 5.86 | 7.37 | 6.49 | 7.69 | 7.06 |
| 12 | 3.01 | 2.88 | 3.08 | 3.51 | 3.19 | 3.33 |
| 13 | 9.20 | 7.01 | 9.30 | 9.05 | 9.89 | 10.38 |
| 14 | 4.39 | 3.81 | 4.84 | 4.08 | 7.06 | 7.76 |
| 15 | 9.11 | 8.32 | 9.33 | 8.23 | 10.09 | 9.48 |
| 16 | 3.88 | 3.55 | 4.24 | 3.87 | 5.75 | 5.54 |
| 17 | 4.47 | 3.41 | 5.00 | 4.48 | 8.23 | 8.00 |
| 18 | 6066.00 | 687.30 | 6082.00 | 692.50 | 6119.00 | 646.70 |
| 19 | 57.30 | 42.60 | 57.48 | 51.80 | 57.83 | 42.60 |
| 20 | 8.56 | 8.85 | 8.58 | 7.99 | 8.63 | 8.94 |
| 21 | 13.64 | 15.97 | 13.66 | 14.40 | 13.72 | 14.56 |
| 22 | 2.86 | 2.88 | 2.95 | 3.04 | 3.23 | 3.33 |
| 23 | 7.86 | 8.41 | 8.02 | 6.94 | 8.56 | 8.37 |
| 24 | 7.04 | 5.54 | 7.08 | 7.28 | 7.19 | 8.00 |

Table IV

COMPUTED AND SIMULATED MEAN CORE LENGTH

| CASE NO. | MEAN CORE LENGTH (WORDS) | | | | | | | PERCENT DEVIATION $100\ (L_c - L_s)/L_c$ | |
|---|---|---|---|---|---|---|---|---|---|
| | Computed | | | | | | Simulated $L_s$ | Rule 1 | Rule 2 |
| | According to Rule 1 | | | According to Rule 2 | | | | | |
| | $L_{Q1}$ | $L_{Q2}$ | Total—$L_c$ | $L_{Q1}$ | $L_{Q2}$ | Total—$L_c$ | | | |
| 1 | 460 | 680 | 1,140 | 465 | 695 | 1,160 | 651 | 42.9 | 43.9 |
| 2 | 1030 | 1,610 | 2,640 | 1035 | 1,625 | 2,660 | 1,700 | 35.6 | 36.1 |
| 3 | 520 | 1,790 | 2,310 | 535 | 1,845 | 2,380 | 3,400 | -47.2 | -42.7 |
| 4 | 1170 | 4,110 | 5,280 | 1180 | 4,170 | 5,350 | 7,500 | -42.0 | -40.2 |
| 5 | 1160 | 1,470 | 2,630 | 1170 | 1,475 | 3,645 | 1,900 | 27.8 | 28.2 |
| 6 | 1010 | 3,370 | 4,380 | 1020 | 3,370 | 4,390 | 5,100 | -16.4 | -16.3 |
| 7 | 990 | 2,360 | 3,350 | 995 | 2,385 | 3,380 | 3,700 | -10.4 | - 9.4 |
| 8 | 950 | 1,460 | 2,410 | 960 | 1,465 | 2,425 | 1,400 | 42.0 | 42.2 |
| 9 | 970 | 3,370 | 4,340 | 980 | 3,380 | 4,360 | 5,800 | -33.7 | -33.0 |
| 10 | 960 | 2,330 | 3,290 | 965 | 2,335 | 3,300 | 3,500 | - 6.4 | - 6.1 |
| 11 | 545 | 1,250 | 1,795 | 565 | 1,300 | 1,865 | 2,000 | -11.4 | - 7.3 |
| 12 | 505 | 715 | 1,220 | 520 | 740 | 1,260 | 1,000 | 18.0 | 20.5 |
| 13 | 595 | 1,530 | 2,125 | 620 | 1,605 | 2,225 | 3,000 | -39.3 | -34.9 |
| 14 | 1010 | 745 | 1,755 | 1060 | 810 | 1,870 | 1,700 | 3.1 | 9.0 |
| 15 | 670 | 1,520 | 2,190 | 700 | 1,610 | 2,310 | 2,900 | -32.4 | -25.4 |
| 16 | 735 | 665 | 1,400 | 780 | 710 | 1,490 | 1,200 | 14.3 | 19.5 |
| 17 | 965 | 690 | 1,655 | 1040 | 755 | 1,795 | 1,700 | - 2.7 | 5.3 |
| 18 | 1060 | 1,342,000 | 1,343,000 | 1075 | 1,356,000 | 1,357,100 | 613,000 | 52.9 | 54.8 |
| 19 | 1050 | 12,880 | 13,930 | 1060 | 13,010 | 14,070 | 25,000 | -79.5 | -77.7 |
| 20 | 970 | 1,805 | 2,775 | 975 | 1,815 | 2,790 | 2,200 | 20.7 | 21.2 |
| 21 | 950 | 1,960 | 2,910 | 955 | 1,965 | 2,920 | 2,600 | 10.7 | 11.1 |
| 22 | 1320 | 1,930 | 3,250 | 1335 | 1,970 | 3,305 | 3,500 | - 7.7 | - 5.9 |
| 23 | 605 | 1,315 | 1,920 | 635 | 1,380 | 2,015 | 2,400 | -77.1 | -68.8 |
| 24 | 480 | 1,295 | 1,775 | 495 | 1,330 | 1,825 | 2,300 | -29.6 | -26.3 |

of new requests and return requests is $\lambda + 1/h_d$. It follows that the computer utility due to the analysis of requests is:

$$\rho_c' = (\lambda + 1/h_d)h_c \quad . \tag{9}$$

As may be seen from Eqs. (A-3) and (A-4), under the assumption of $\rho_c = 1$, which is the worst case, the ratio of the computer utility due to interruption $\rho_c''$ and the computer utilization due to the analysis of requests $\rho_c'$ is:

$$\frac{\rho_c''}{\rho_c'} = \frac{h_h[1 + (N - 1)Z]}{h_c N} \quad . \tag{10}$$

For the system under study, $X = 1/4$, $Z = 0.76$, and $h_h = 1$ msec; for Case 18 where $h_c = 50$ msec, the ratio $\rho_c''/p_c' = 0.015$, which means $\rho_c''$ is small in comparison to $\rho_c'$. Thus, $\rho_c$ may be approximated by $\rho_c'$. More specifically, under disc saturation conditions, $\rho_c$ may be approximated by Eq. (9). For Case 18, where $\lambda = 1/3$ requests/sec and $h_d = 0.36$ sec/request, $\rho_c' = \rho_c = 15.55\%$, which compares favorably with the simulated result of 15.31%. Thus, under disc saturation conditions Eq. (9) instead of Eq. (1) should be used to compute the utility of the computer.

The computed $T_r$ for Case 18 is about nine times the simulated $T_r$; for Case 19, the computed $T_r$ is about 16% higher than the simulated $T_r$. In both of these cases, the disc utility is near or at 100%. It is assumed the simulation started from the empty state, with no request in the system, and then proceeded to build up to its steady state level. In the case of 100% utility there is no steady state level since the queue at the disc will continue to build up without bound. The time of buildup could be rather long when the utility of the server is near 100%. The simulated results are based on 1000 reply messages being generated within the system and sent out; therefore, it is conjectured that the simulated $T_r$ for Cases 18 and 19 does not represent the steady state $T_r$. They are heavily influenced, especially the $T_r$ of Case 18, by the initial condition of the system. It is concluded that either a longer simulation run is needed for situations like Cases 18 and 19, or that means should be devised to eliminate the bias caused by the transient condition.

The computed $\rho_c$ are less than the simulated $\rho_c$ in all cases, with the single exception of Case 18, as was noted above. One possible explanation is the bias introduced in the simulated mean arrival rate and service time, as evident in the bias introduced in the mean disc search. In all cases, the simulated mean disc search is less than the specified mean disc search.

The computed mean response times for each of the three priority classes seems to be in general agreement with the simulated results (see Table III), except in the Cases 18 and 19. The above remarks on these two cases apply here too.

It is to be noted in Cases 3, 4, 7, 19, 20, 21, and 23 that the simulated mean response time for the higher priority messages is greater than that of the lower priority messages. This inconsistency is probably due to sampling errors of the simulation process.

The simulated mean core length for each case, as shown in Table IV, represents a single sample observation. Therefore, it is rather difficult to indicate the degree of agreement between the analytical models and the simulated model. However, an observation may be made on the percentage deviation of the simulated mean from the analytical mean. Table V shows the proportion of cases that deviate less than $x$ percent for Storage Rule 1 and 2. Also shown is the normal distribution with $\sigma = 1/3$ mean.

Table V

CUMULATIVE DISTRIBUTION OF PERCENT DEVIATION
OF SIMULATED MEAN CORE LENGTH
FROM THE ANALYTICAL MEAN

| PERCENT DEVIATION | PROPORTION OF CASES | | NORMAL DISTRIBUTION |
|---|---|---|---|
| | Rule 1 | Rule 2 | |
| $\leq 10$ | 16.7 | 25.0 | 23.6 |
| $\leq 20$ | 41.7 | 41.7 | 45.1 |
| $\leq 30$ | 54.1 | 62.5 | 63.2 |
| $\leq 40$ | 70.9 | 83.4 | 77.0 |
| $\leq 50$ | 87.5 | 91.7 | 86.6 |
| $\leq 60$ | 91.7 | 96.0 | 92.8 |
| $\leq 70$ | 91.7 | 96.0 | 96.4 |
| $\leq 80$ | 100.0 | 100.0 | 98.4 |

A comparison of the observed percent deviation distribution with the normal distribution seems to indicate the simulated means have a normal distribution in relation to the analytical mean. This implies a reasonable agreement exists between the analytical means and the simulated means.

## E. CONCLUSIONS

The analytical analysis of the 473-L System indicates it is possible to apply elementary queuing theory to analytically model a relatively complex information system, and to obtain some of the queuing measures of the system. The validity of the analytical model may be verified by simulation. Once verified, the analytical model may be used by the system designer in considering the trade-off of system elements.

## III NUMERICAL EVALUATION OF WAITING-TIME DISTRIBUTION

### A. INTRODUCTION

As seen in Sec. II, by suitable idealization of the information system we can apply the queuing theory to obtain the means of the queuing measures of the system. The information system designer is often interested in the distributions of the queuing measures so that end point conditions of the system can be analyzed. However, analytical expressions for the distributions of priority queuing measures usually are not available, or if available they seldom are in readily usable form. Thus, an approximate analytical expression that is readily computable is a welcome addition to the information system design.

During the previous research effort, Eq. (8) of Reference 2, the waiting-time distribution of the second-priority customer, $W_2(T)$, in an exponential head-of-the-line priority queuing system has been shown to be

$$W_2''(>T) = \frac{\rho^2 - \rho_1}{\rho - \rho_1} e^{-\alpha T} + \frac{1-\rho}{2\pi} \int_{\alpha_1}^{\alpha_2} f(r)dr, \qquad \text{for} \qquad \rho^2 \geq \rho_1 \quad . \tag{11a}$$

$$= \frac{1-\rho}{2\pi} \int_{\alpha_1}^{\alpha_2} f(r)dr, \qquad \text{for} \qquad \rho^2 < \rho_1 \quad , \tag{11b}$$

where

$\rho$ = the loading factor of the system

$\rho_1$ = the loading factor of the system due to the first priority customers

$T$ = unit of time in units of mean service time, $\mu$.

$$\alpha_1 = (1 - \sqrt{\rho_1})^2$$

$$\alpha_2 = (1 + \sqrt{\rho_1})^2$$

$$\alpha = \frac{(\rho - \rho_1)(1 - \rho)}{\rho}$$

$$f(r) = \left[ e^{-rT} \frac{\sqrt{(\alpha_2 - r)(r - \alpha_1)}}{[r(r - \alpha)]} \right] .$$

Cox[3] has derived a simple expression for the distribution, namely

$$W_2''(>T) = \rho e^{-(1-\rho_1)(1-\rho)T} . \qquad (12)$$

But Cox's expression is not valid for all ranges of $\rho_1$ value. Since Eq. (12) is a relatively simple equation it is of interest to investigate the range of $\rho_1$ value where $W_2''(>T)$ is valid.

It is seen that when $\rho_1 << \rho$, then $(\rho^2 - \rho_1)/\rho - \rho_1) \to \rho, \alpha \to 1 - \rho$ and $\int_{\alpha_1}^{\alpha_2} f(r)dr \to 0$. Therefore, Eq. (11) becomes that of Eq. (12). It is also to be noted that when $\rho << 1$ implies $\rho_1 << 1$, which means $\int_{\alpha_1}^{\alpha_2} f(r)dr \to 0$. Thus, Eq. (11) again tends toward Eq. (12).

## B. RESULTS

A numerical evaluation of Eq. (11) was carried out for $\rho$ = 0.1, 0.2, 0.3, 0.5, 0.7, 0.85, and 0.95, and for $\rho_1/\rho$ = 0.1, 0.2, 0.5, and 0.9. The results of the evaluation are plotted in Fig. III-1 through 6 for each value of $\rho$ except $\rho$ = 0.1. When $\rho$ = 0.1 Eq. (11) and Eq. (12) are almost identical. The abscissa of the figures is in units of $X = [20(1 - \rho)(1 - \rho_1)T]/(\ln \rho - \ln 10^{-2})$. This unit of $X$ is chosen so that Eq. (12) appears as single curve for a specific value of $\rho_1$. In the figures the unmarked straight line is the waiting-time distribution due to Cox, $W_2''(>T)$.

We see from the figures that the waiting-time distribution due to Cox approximates that of Eq. (11) quite well for $\rho_1 \leq 0.2\rho$ and for all values of $\rho$ with the possible exception when $\rho$ is near 0.5. When $\rho$ is near 0.5, the percentage deviation of $W_2'(>T)$ from $W_2''(>T)$ becomes relatively large at the extreme value of $X$, as much as 45% at $X$ = 20. As the ratio of $\rho_1/\rho$ increases, the percentage deviation becomes larger, which verified the analytical observation made above.

From the figures we note that Eq. (12) yields a more optimistic estimate of waiting-time at the lower value of $X$ and a more pessimistic estimate at higher value of $X$. The cross-over point, where $W_2'(>T) = W_2''(>T)$, varies with the value of $\rho$.

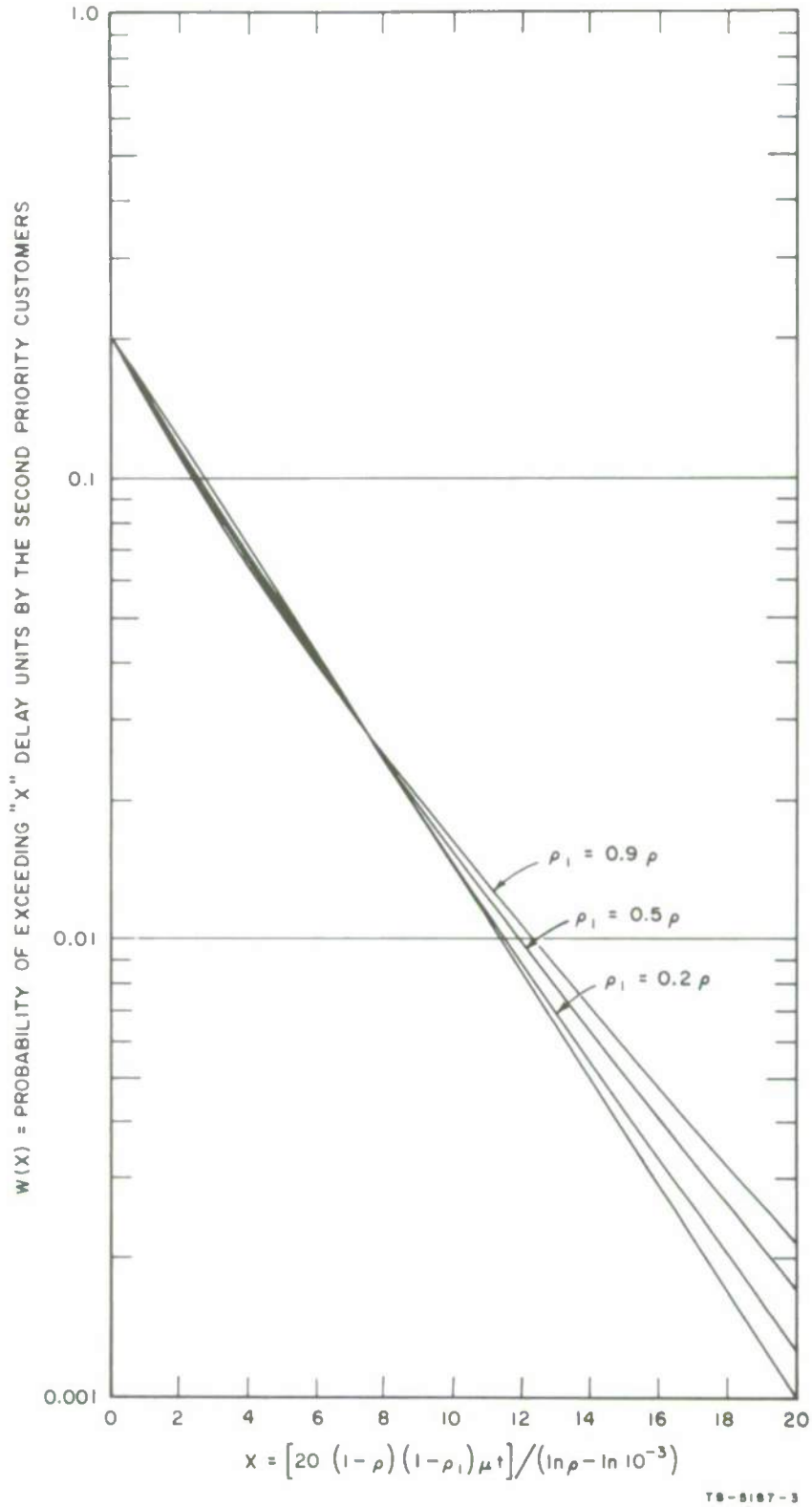



FIG. III-1 WAITING-TIME DISTRIBUTION OF THE SECOND PRIORITY — $\rho = 0.2$

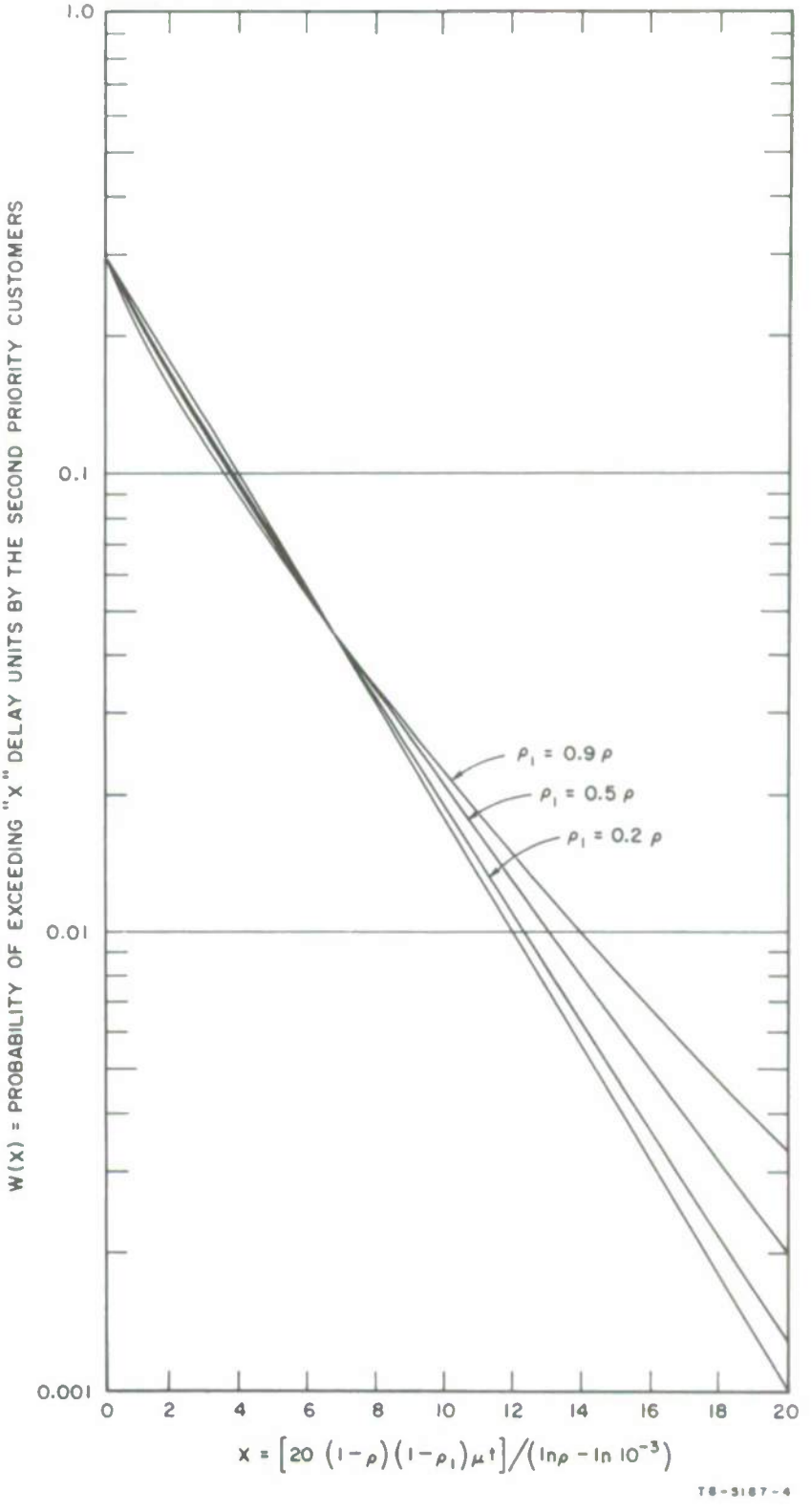



FIG. III-2 WAITING-TIME DISTRIBUTION OF THE SECOND PRIORITY — $\rho = 0.3$

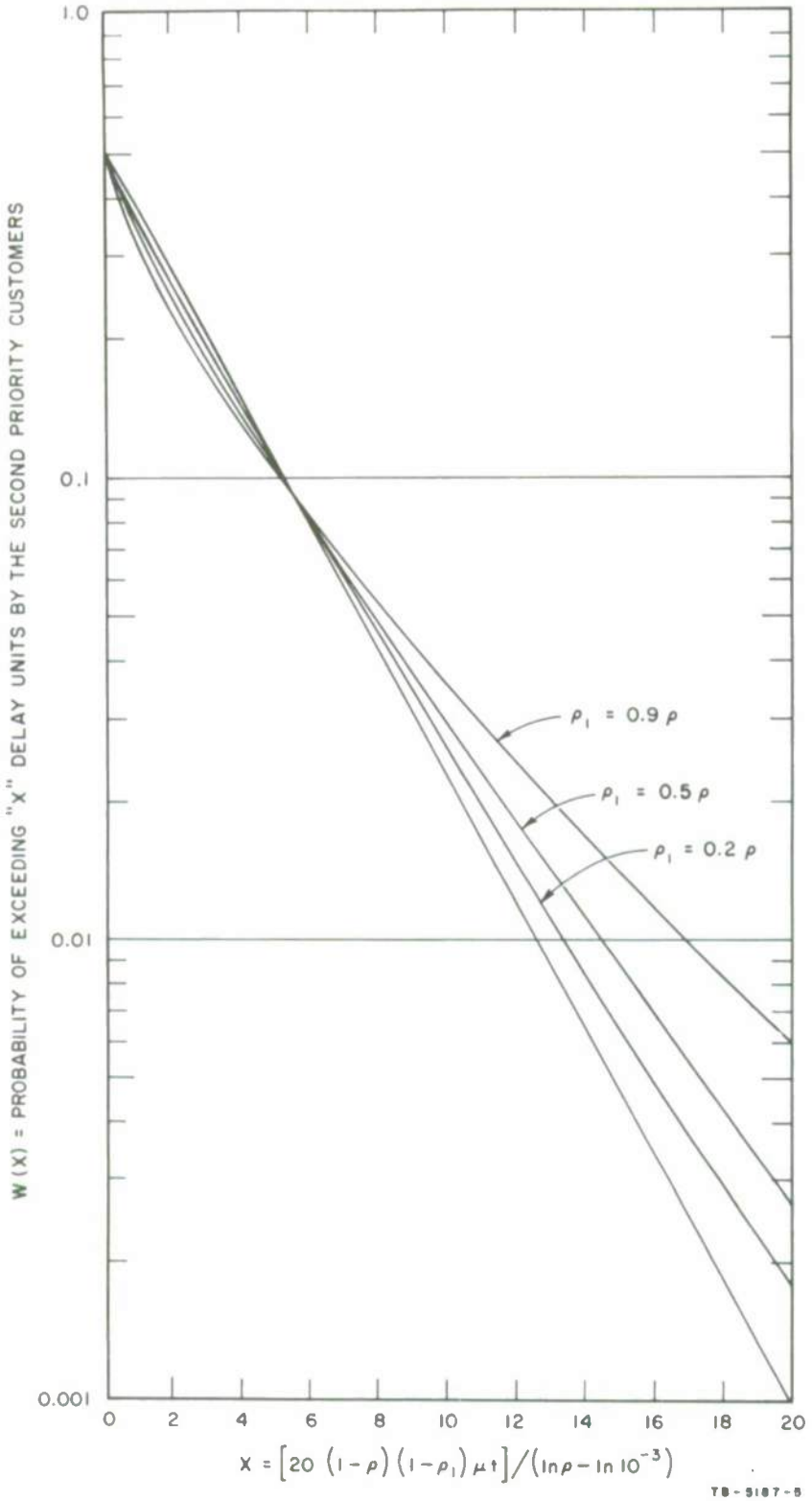



FIG. III-3 WAITING-TIME DISTRIBUTION OF THE SECOND PRIORITY — $\rho = 0.5$

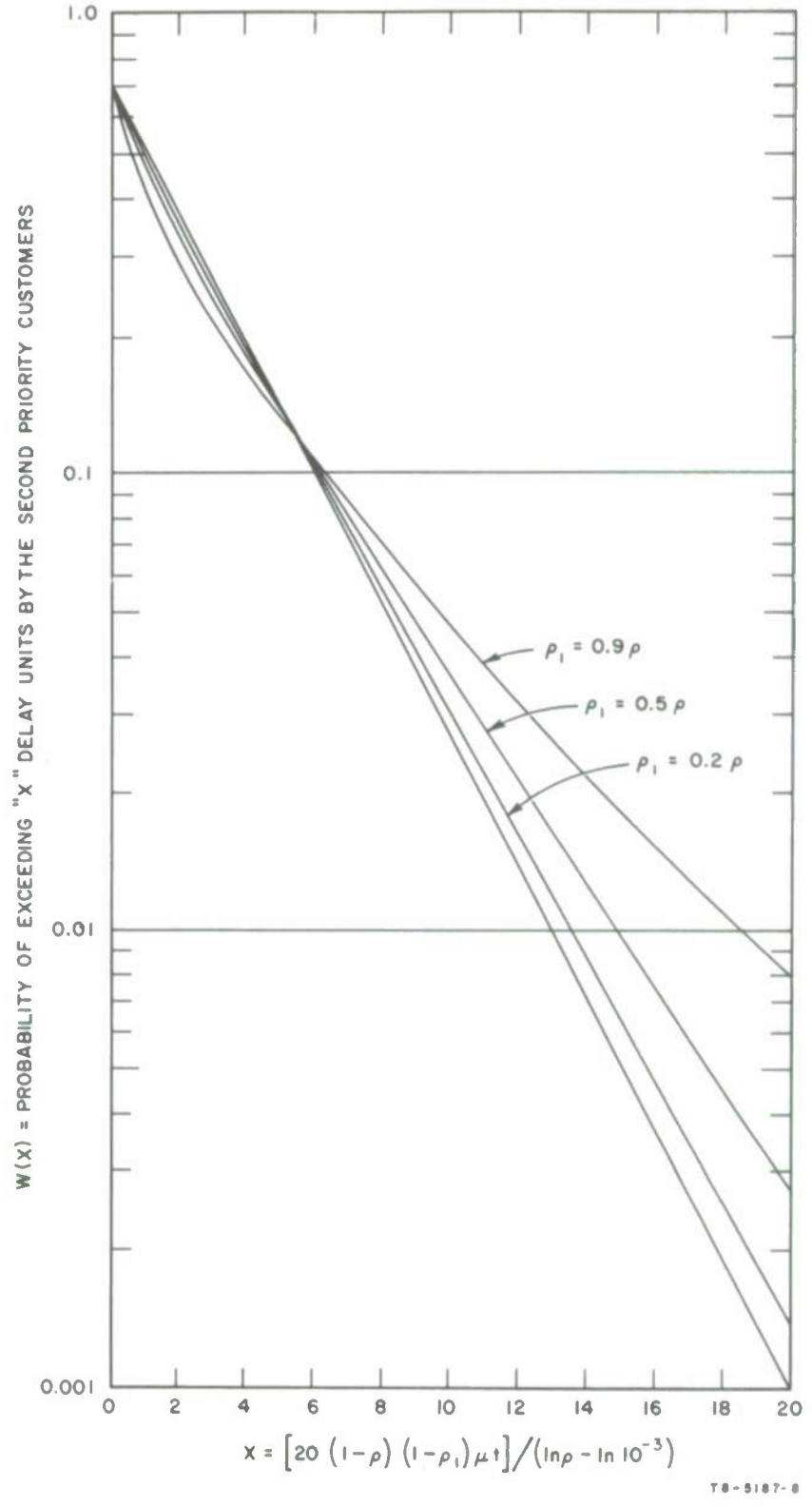



FIG. III-4 WAITING-TIME DISTRIBUTION OF THE SECOND PRIORITY — $\rho = 0.7$

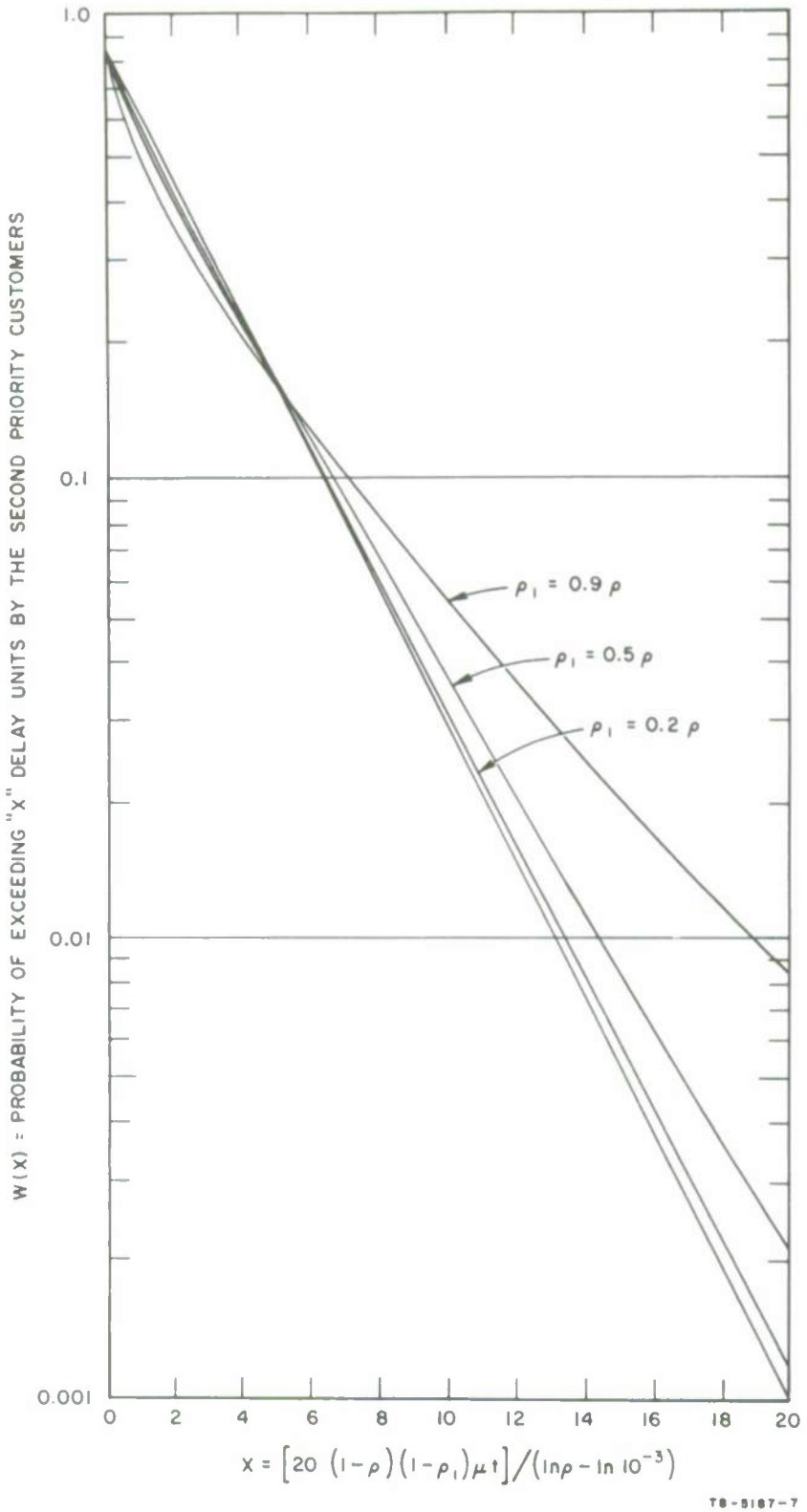



FIG. III-5 WAITING-TIME DISTRIBUTION OF THE SECOND PRIORITY — $\rho = 0.85$

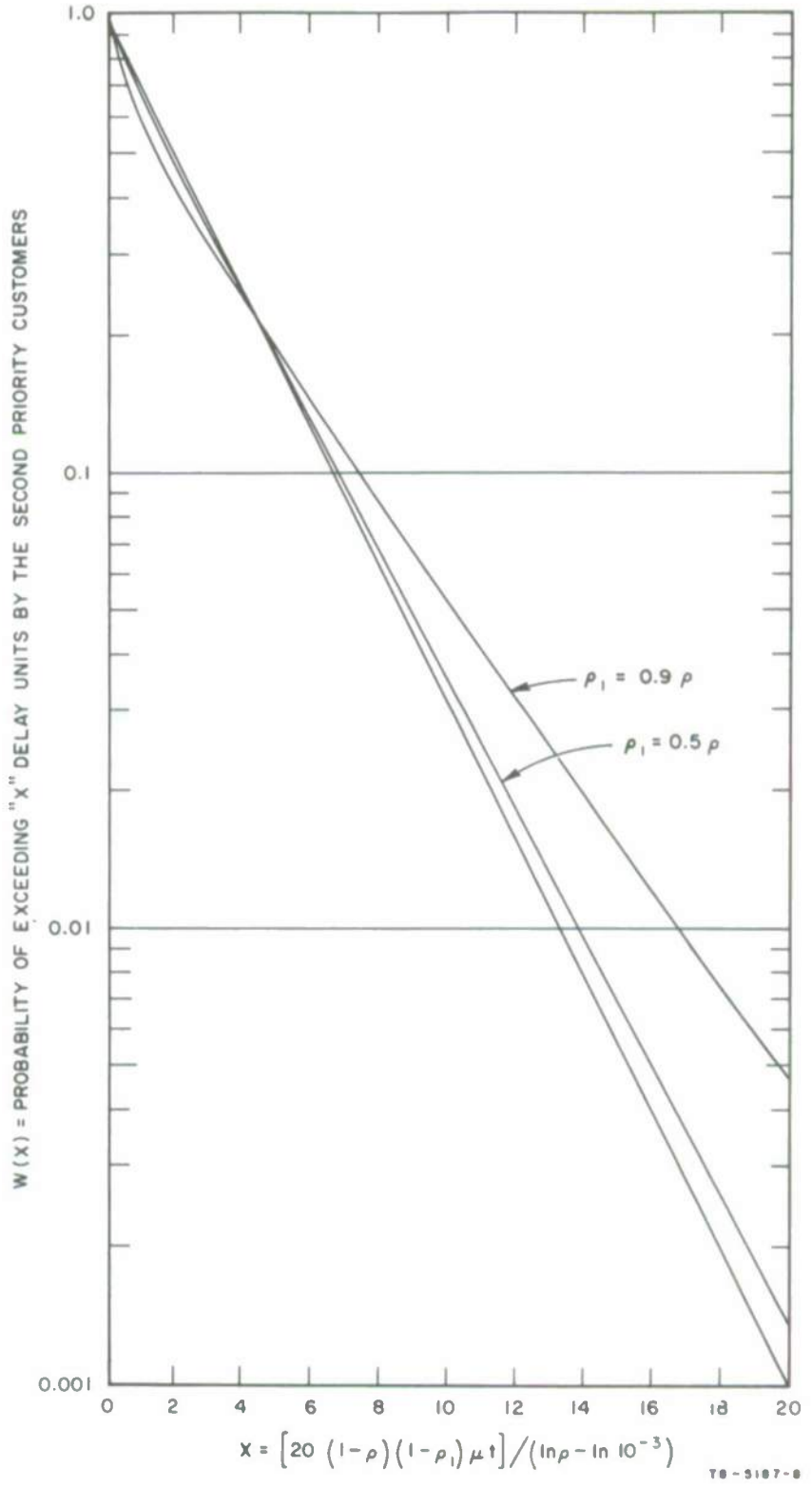



FIG. III-6 WAITING-TIME DISTRIBUTION OF THE SECOND PRIORITY — $\rho = 0.95$

# IV SIMULATION STRATEGY

## A. INTRODUCTION

In many information systems the customers (messages) are categorized into priority classes and their order of service is in accordance with a given priority queuing discipline. An example of this type of information system is the 473-L system, as described in Sec. II. It is sometimes possible to idealize the parameters of the system and analyze its queuing measures analytically, as was done in Sec. II. Because there is a relative scarcity of analytical results on the priority queue,[2] however, it is not always possible to analyze the queuing measures without extreme idealizations. Such extreme idealizations may make the results of the analysis unrealistic. Under such a circumstance the method of simulation frequently is used to analyze the queuing measures or to check the validity of idealizations. It would seem advisable, therefore, to develop a guide for efficient design of the simulation experiment.

The present study investigated the stopping rules and the chopping rules of the simulation experiment, and the application of "Variance Reduction Methods" to the Simulation of Priority Queuing Systems.

## B. STOPPING RULES

In simulation the question often asked is how long should be the simulation experiment? Alternately the question is when to stop the simulation experiment? One possible answer is to stop the simulation experiment when the observed variation in the estimates reaches certain specified confidence interval sizes. Based upon a concept of Conway,[4] the confidence interval about $\overline{X}$ has been shown to be (Appendix B):

$$\overline{X} \pm \alpha S(\overline{X}_n)$$

where $\alpha$ = the number of standard deviations in an $(1 - \alpha)\%$ confidence interval.

$$S^2(\bar{X}) = \frac{S^2(X_n)}{n} \frac{1 + \rho_E}{1 - \rho_E}$$

$$S^2(X_n) = \frac{1}{n-1}\left[\sum_{t=1}^{n} X_t^2 - n(\bar{X}_n)^2\right]$$

$$\bar{X}_n = \frac{1}{n}\sum_{t=1}^{n} X_t$$

$$X_t = \frac{1}{L}\sum_{j=1}^{L} Y_{tj}$$

$Y_{tj}$ = the $j$th observation of the queuing measure in the $t$th block of the simulation experiment. (13)

$$\rho_E = \rho + \frac{n+2}{(n-1)^2} \frac{1+\rho}{1-\rho}$$

$$\rho = C/[(n-1)S^2(X_n)]$$

$$C = \sum_{t=1}^{n-1} (X_t X_{t+1}) - (n-1)(\bar{X}_{n1})(\bar{X}_{n2})$$

$$\bar{X}_{n1} = \frac{1}{n-1}\sum_{t=1}^{n} X_t$$

$$\bar{X}_{n2} = \frac{1}{n-1}\sum_{t=1}^{n-1} X_t \quad . \qquad (14)$$

Given $L$ and $\alpha$, it is possible to determine the confidence interval about $\bar{X}_n$ at any value of $n$, where $n$ is the length of simulation. Since the confidence interval can be computed as the simulation progresses, it is possible to stop the simulation whenever the confidence interval is less than certain specified size.

Any stopping rule which depends on the observations themselves can prodnce bias. For example, suppose that the sample mean $\overline{X}_n$, and sample variance $S^2(\overline{X}_n)$ are independently distributed. This is true if $X_t$ is normally distributed. Suppose the stopping rule is:

$$\text{Stop when} \qquad S(\overline{X}_n) \leq K\overline{X}_n$$

Then when, for a given $S(\overline{X}_n)$, $\overline{X}_n$ is low the simulation will be stopped. If $\overline{X}_n$ is high, further samples will be taken which will tend to reduce $\overline{X}_n$. Thus this rule favors lower values of $\overline{X}_n$ and is biased.

For $X_t$ normally distributed, the rule:

$$\text{Stop when} \qquad S(\overline{X}_n) \leq K$$

is unbiased because now the stopping point is independent of $\overline{X}_n$.

Another example is $X_t$ with an exponential distribution. Here $S(\overline{X}_n)$ tends to be proportional to $\overline{X}_n$. $S(\overline{X}_n) \approx \alpha\overline{X}_n$. Then the rule:

$$\text{Stop when} \qquad S(\overline{X}_n) \leq K$$

is biased because $S(\overline{X}_n) \leq K$ implies $\overline{X}_n \leq K/\alpha$ and again low values of $\overline{X}_n$ will satisfy the rule and stop simulation. In this case the rule:

$$\text{Stop when} \qquad S(\overline{X}_n) \leq \overline{K}X_n$$

will be approximately unbiased.

The $X_t$ can usually be assumed to be normally distributed, therefore the unbiased stopping rule should be:

$$\text{Stop when} \qquad S(\overline{X}_n) = K \quad .$$

## C. CHOPPING RULES

A problem in the simulation of queuing systems is the selection of initial state. If the initial state is not selected at random from the true distribution of states when the system is in steady-state the choice will introduce some bias into the resultant average statistics. A

possible course of action to avoid bias is to start the simulation from empty state and chop-off the first part of the run from the average statistics.

A chopping rule was developed and is presented in Appendix C. This chopping rule is based upon the assumption that the simulation experiment would start from the empty state and progress to the steady-state. Hence the parameter, which in this case would be the mean waiting-time as estimated from the initial part of the simulation, would be lower than that obtained from the later part of the simulation. A statistical test procedure that can test the hypothesis that the parameter as estimated from the initial part of the simulation is less than that from the later part of the simulation can be used to determine the point of chopping. The testing procedure is as follows (using the notations of the stopping rule):

Let the sequence of the estimates of the parameter from the simulation be

$$\bar{X}_m = \frac{1}{m} \sum_{t=4}^{n} X_t, \quad m = 1, 2, 3, \ldots, n \quad .$$

Starting from $m = 1$ compute the statistics

$$t_m = \frac{\bar{X}_{n-m} - \bar{X}_m}{\left(\frac{1}{n-m} + \frac{1}{m}\right)^{\frac{1}{2}} \left[\frac{(m-1)S^2(X_m) + (n-m-1)S^2(X_{n-m})}{n-2}\right]^{\frac{1}{2}} \left(\frac{1+\rho_E}{1-\rho_E}\right)^{\frac{1}{2}}} \quad (15)$$

where

$$\bar{X}_{n-m} = \frac{n\bar{X}_n - m\bar{X}_m}{n-m}$$

$$S^2(X_{n-m}) = \frac{1}{n-m-1} \left[(n-1)S^2(X_n) - (m-1)S^2(X_m) - (n-m)\bar{X}^2_{n-m} + n\bar{X}^2_n - m\bar{X}^2_m\right] \quad .$$

Find $m' = \max(m : tm\ 2)$ and chop at $m'$.

Based upon experience one may ignore the chopping rule unless the system loading factor is greater than 0.9.

## D. THE APPLICATION OF "VARIANCE REDUCTION METHODS" TO THE SIMULATION OF PRIORITY QUEUING SYSTEMS

### 1. General

#### *a*. Introduction

The history of variance reduction is very closely associated with that of stochastic simulation in general, and it is natural they are often jointly referred to by the term "Monte Carlo."* The queuing problem is mentioned in many papers as a natural example for this method, and is dealt with specifically in References 5, 6, and 7. Only a few of the ideas suggested therein, however, can be applied directly to the priority queuing situation. Some may introduce bias, others lose most of their efficiency. The problem of the distribution of the initial position increases in dimensionality and becomes almost too formidable to tackle. On the other hand, there are some advantages in these methods not available in the non-priority case. Often, there is at least one estimator more than there are estimands, which suggests the use of a regression model. Also, in some cases the difficulty is only with the priority discipline. The overall mean waiting time (or mean waiting time for some of the priority classes) is known analytically, and this knowledge can be incorporated in the analysis of the experiment to reduce the variance of the estimator.

The utility of simulation depends not only on getting estimators with sufficiently small variances, but also on the ability to accurately estimate these variances. As has been pointed out by Kahn and Marshall, "nothing is worse than thinking you have a good estimate when in fact you have a very bad one."[8] The problem therefore, is twofold: to design the experimental setup to yield practically no bias in the estimates and their confidence interval, and to reduce the variance of a fixed volume of sampling without a similar increase in computation labor.

#### *b*. Experimental set up of the simulation

The general setup of the experiment followed closely the discussion in Conway.[4] The duration of the run was determined by the time until the departure of a fixed number of arrivals, *i.e.*, "run until the

---

* An up-to-date survey of the development of Monte Carlo methods is presented in Reference 5, where an extensive bibliography is also to be found.

first 5000 arrivals have departed." Customers were grouped into sets of fixed size, according to the time of their arrival. Overall mean waiting time for customers in a set and the means corresponding to the varions priority classes as well as additional input statistics of the set were computed for each of the sets, printed out, and stored for the final summary. Thus, the individual sets are considered as separate simulations, the initial point of which is the final position of its predecessor, and therefore the sets are properly drawn at random from the distribution corresponding to the simulated load.

Experiments were started from the "empty and idle" position. This avoids the difficulties involved in choosing from a distribution of high dimensionality, and therefore one that would be difficult to estimate in a preliminary run. Beginning in this manner, however, introduces bias, and usually requires that the first few observations be chopped off and not considered in the summary analysis. Theoretically, the bias introduced by an arbitrary initial point (any point) can never be totally eliminated, as indeed the "steady state" can never be reached in a simulation run, no matter how long. In practice, however, even mere chopping was found unnecessary for the simulation lengths used with most of the loading factors analyzed. With loadings of over 0.9 Erlang, this problem influences the estimator more strongly. Statistical tests of the transient effect with subsequent chopping are necessary, as presented in Appendix C of this report. Throughout the present report the transient effect is disregarded.

It would be proper at this point to define the general setup and basic output:

$\text{Cust}_n$: = $n$th arrival

$\text{Set}_t$: = $\{\text{Cust}_n : n = (t-1)m + 1, \ldots, tm\}$ $\quad t = 1, \ldots, T$

$C_p$ = priority class $p$ (all customers with priority $p$)
$p = 1, \ldots, NP$ where $p = 1$ is the highest priority

$NP$ = number of priority classes

$\text{Wait}(n)$ = Waiting time of $\text{Cust}_n$

$\text{Set}_{tp} = \text{Set}_t \bigcap C_p$ = all customers in $\text{Set}_t$ that have priority $p = 1, \ldots, NP$

$N_{tp}$ = number of priority $p$ customers in set, $p = 1, \ldots, NP$

where

$$r = \frac{\text{Cov } [X_t, X_{t+1}]}{\text{Var } X_t} \quad \text{and} \quad r^k = \frac{\text{Cov } [X_t, X_{t+k}]}{\text{Var } X_t} \quad , \quad r < 1$$

with the inequality approaching equality for large $m$.

This method involves at least two substantial difficulties: first $S_x^2$ is no longer an unbiased estimator of $\sigma_x^2$ but

$$ES_x^2 < \sigma_x^2 \quad \text{for} \quad r > 0 \quad .$$

Second, the estimate of Var $\bar{X}$ depends on a good estimate of $r$, and when $r$ is large even a small variance in the estimate of $r$ will induce a large variance in the estimate of Var $(\bar{X})$. Hence, even if independence is not to be assumed, $m$ has to be large enough to keep $r$ small. Also, the same data may be arranged as sequences with different values of $m$ (and $T$) and the estimates compared.

In most cases, however, these elaborate measures are not justified, and one may simply impose lower bounds on $m$ and $T$, with the implication that the volume of the experiment may be determined not necessarily by the variance of the estimates, but possibly by the variance of the estimate of these variances.

In the simulation of priority queues, we get several estimators simultaneously. The volume of the experiment is determined by the "worst" of these estimators, in terms of the accuracy required. Thus, although usually the class with lowest priority will have the highest variance, it may be that only a rough estimate is required for that class. The length of the experiment would be determined by a higher priority class with a smaller variance. This is especially true since accuracies are usually specified as a percentage of the estimated value, and for the lowest priority not only the variance but also the estimand itself is highest.

Adaptive rules to determine the length of the simulation may be developed, as discussed in Appendix B of this report. The stopping rule tends to be biased—sometimes as to the estimands themselves, but more often with respect to the estimate of the variance. Possibly, the

best course is to extrapolate the required volume from a short preliminary run. Note, however, whenever there is a rigid requirement on the estimate of the variance (*i.e.*, the results have to be presented as $P[0.95\ W \leq E\ [X] \leq 1.05\ W] \geq 0.95$), there is a tendency to overestimate the confidence. This is avoided if a certain variability in the stated confidence is allowed.

## 2. Variance Reduction

### *a.* Introduction

The basic principle of variance reduction methods involves taking advantage of analytical knowledge of part of the simulated process. By taking account of the correlation between input and output, one can account for part of the variance of the output by a proper analysis of the input, or eliminate part of that variance by a proper manipulation of the input. To be done effectively, this requires insight into the probabilistic structure of the simulated process. This is the reason why "Monte Carlo" often seems a collection of "ad-hoc" methods. This is also why the application of established principles to a specific problem is not trivial. Even in the limited field of priority queues, different methods will have different effectiveness for different loadings.

The measure of efficiency of a certain design relative to direct simulation may be expressed as a ratio $F$, according to Hammersley and Handcomb,[5] where

$$\text{"Efficiency ratio"}\ F = \frac{n_1\sigma_1^2}{n\sigma^2} ,$$

$$\text{"Labor ratio"}\ L = \frac{n_1}{n} ,$$

$$\text{"Variance ratio"}\ \psi = \frac{\sigma_1}{\sigma} ,$$

where $n_1$ and $n$ represent the computational volumes required, and where $\sigma_1^2$ and $\sigma^2$ represent the variances of the estimate for the investigated design and direct simulation, respectively. Whenever there is a set of estimands, $\psi$ should represent the ratio for the "worst" estimate, in the sense discussed. Since under different designs the worst estimate may relate to different estimands, it is sometimes preferable to consider the efficiency as

$$V = \frac{n_1^1}{n^1}$$

where $n_1^1$ and $n^1$ are the "required" volumes of computation for the specified accuracy.

An alternative measure suggested by Ehrenfeld and Ben-Tuvia,[9] is the relative reduction in variance

$$\eta = \frac{\sigma^2 - \sigma_1^2}{\sigma^2}$$

for a fixed volume of computation. This is expressed as a fraction or percentage. Both measures are used in this report, and they are either directly specified or clear from the context.

### *b.* THE PRIORITY QUEUE AS A STOCHASTIC PROCESS

Considering a set of size $m$ as an independent run, the stochastic process ( and subsequently the output) is defined by a set of random variables. The initial point is a random vector, which in the case of priority disciplines has a high dimensionality. Even in the simplest case of Poisson arrivals and exponentially distributed service time the dimension of the initial point is $NP$. When $m$ is large, however, variation of the output due to the initial point is fortunately small, and it is not necessary to attempt to reduce this part of the variance.

Given the initial point, the run is determined by a sequence of random interarrival times and service times for each of the priorities. The values are drawn from their respective distributions. Alternatively, this can be viewed as a single sequence of random numbers representing interarrival times, service times, and the priority class. Variation in waiting times, and subsequently in mean waiting times, is caused by variations in the input sequence. These may be classified as variations in magnitudes and variations in the permutations. The magnitudes are easier to classify, analyze and control, but it is the permutations that account for a larger part of the variance.

*c.* CONTROL VARIATES

The basic idea is as follows: if $X$ is an unbiased estimate of $E[X]$, then

$$W = X - \alpha\tilde{V} \qquad \left(\text{where } \tilde{V} = V - E[X]\right.$$
$$\left.V = \text{control variate, and } \alpha \text{ is a constant.}\right)$$

is also an unbiased estimator. We have

$$\text{Var } W = \text{Var } X + \alpha^2 \text{ Var } V - 2\alpha \text{ Cov } [X,V] \quad .$$

To find optimal value of $\alpha$, differentiate

$$\frac{\partial}{\partial\alpha} \text{Var } W = 2\alpha \text{ Var } V - 2 \text{ Cov } [X,V] = 0 \quad ,$$

$$\rightarrow \alpha = \frac{\text{Cov } [X,V]}{\text{Var } V} \quad ,$$

$$\rightarrow \text{Var } W^* = \text{Var } X - \frac{\text{Cov}^2 [X,V]}{\text{Var } V} = \text{Var } X(1 - r_{xv}^2) \quad ,$$

and

$$\psi = 1 - r_{xv}^2 \quad , \qquad \eta = r_{xv}^2 \tag{17}$$

where $r_{xv}$ is the normalized correlation coefficient. If Cov $[X,V]$ is unknown, it can be substituted by its estimate $\hat{\sigma}_{x,v}$ $W$ will remain unbiased only if $\hat{\sigma}_{x,v}$ is independent of $X$ and $V$. This cannot generally be assumed, but the estimator is still consistent. With $T$ sufficiently large the bias can be ignored. Alternatively, tests on bias may be introduced (see Appendix B). $\alpha$ can also be estimated from a preliminary experiment, or a previous one with a similar, though not necessarily equal, load.

The variance of the estimator is estimated by

$$\hat{\sigma}_x^2 = \hat{\sigma}_x^2 - \frac{\left(\hat{\sigma}_{xv}\right)^2}{\sigma_v^2} \quad . \tag{18}$$

When $\hat{\sigma}_x^2$ and $(\hat{\sigma}_{xv})^2$ are strongly and positionally correlated, this has a much smaller variance than the variance of $\sigma_x^2$. In our experiments there is a clear indication that this was the case.

A control variate is a known random variable, *i.e.*, $EV$ and Var $V$ are known. The idea can be extended to a vector of control variates (Tocher has presented this idea but develops the $\alpha$ by the least square method).[10]

Let $V$, $E[V]$, $C[V]$ be $k \times 1$ vectors

$$\tilde{V} = V - E(V)$$

$$CV_j = \text{Cov}[X, V_j]$$

Let $\Sigma$ be a $k \times k$ covariance matrix of $V$, $\Sigma_{ij} = \text{Cov}[V_i, V_j]$ .

Let $\Sigma_x = \left[\begin{array}{c|c} \sigma_x^2 & CV \\ \hline CV^t & \Sigma \end{array}\right]$ , which is also a covariance matrix.

$W = x + \alpha\tilde{V}$, where $\alpha$ is a $k \times 1$ vector of coefficients. By solving $k$ simultaneous linear equations of the gradient, we get for optimal $\alpha$

$$\alpha^* = \Sigma^{-1} \cdot CV$$

and

$$\text{Var } W^* = \sigma_x^2 - CV^T \Sigma^{-1} CV = \left|\frac{\Sigma_x}{\Sigma}\right| .$$

It follows from the positive definiteness of the covariance matrix[11] that

$$0 \leq \text{Var } W^* \leq \sigma_x^2 .$$

$\Sigma$ will never be singular, since it obviously would not help us to use control variates which are linear combinations of the others.

At this point it might help to get more insight at the structure of multiple control variates by developing the explicit expression for $k = 2$. In this case,

$$\alpha_1 = \frac{\sigma_x}{\sigma_{v1}} \cdot \frac{r_{x,v1} - r_{x,v2} r_{v1,v2}}{1 - r^2_{v1,v2}}$$

$$\alpha_2 = \frac{\sigma_x}{\sigma_{v2}} \cdot \frac{r_{x,v2} - r_{x,v1} r_{v1,v2}}{1 - r_{v1,v2}}$$

$$\psi = 1 - \eta = 1 - \frac{r^2_{x,v1} + r^2_{x,v2} - 2 r_{x,v1} r_{x,v2} r_{v1,v2}}{1 - r^2_{v1,v2}} \qquad (19)$$

Suppose $V_1$ is the better single control variate ($r^2_{x,x1} > r^2_{x,v2}$). We want to investigate the advantage of introducing $V_2$.

$$\eta_{(v1,v2)} - \eta_{(v1)} = \frac{(r_{x,v2} - r_{x,v1} r_{v1,v2})^2}{1 - r^2_{v1,v2}}$$

We see here that a good "support" for a control variate is not necessarily one that would itself make a good control variate. Suppose, for example, that all three are positively correlated. By assumption, $r_{x,v1} > r_{x,v2} \cdot r_{x,v2}$ may be nearly as high as $r_{x,v1}$, but if $r_{v1,v2}$ is also high there is practically no advantage—*e.g.*,

$$\left.\begin{aligned} r_{x,v1} &= 0.8 \\ r_{x,v2} &= 0.75 \\ r_{v1,v2} &= 0.8 \end{aligned}\right\} \rightarrow \eta_{(v1,v2)} - \eta_{(v1)} = 0.0335 \quad . \qquad (20)$$

A very good situation exists when $r_{x,v2}$ and $r_{x,v1} \cdot r_{v1,v2}$ are of opposite signs, but such control variates are difficult to find. When $r_{v1,v2} = 0$

$$\eta_{(v1,v2)} - \eta_{(v1)} = r^2_{x,v2} \quad ,$$

which would be expected. Another and more surprising possibility is where we have a high $r_{v1,v2}$, and $r_{x,v2} = 0$! Heuristically, this means we improve the estimate by disregarding that part of the control variate fluctuations which have nothing to do with it.

It is clear that the more control variates are operating, the more difficult it would be to find additional ones that would still be effective. But effectiveness should be considered on a relative basis, in terms of variance ratio, and not the efficiency fraction, as, for example, reducing the variance from say 10% to 5% of the original variance involves reducing the necessary volume by half. Namely:

$$\eta_{[v1,\ldots,v(k+1)]} - \eta_{[v1,\ldots,v(k)]} = 0.05$$

$$\frac{\eta_{[v1,\ldots,v(k+1)]} - \eta_{[v1,\ldots,v(k)]}}{1 - \eta_{[v1,\ldots,v(k+1)]}} = 0.5 \qquad (21)$$

Thus, control variates that would initially be regarded as ineffective may prove useful after the important control variates have taken effect. This is because they either serve as good complements, or because they account for a different part of the variation. It seems proper to call them descriptively as "marginal" control variates.

In searching for good control variates in the simulation of queuing systems, the sources of variation should be kept in mind. The problem is to reduce the dimensionality by finding, or composing, a few control variates that would maintain most of the relevant features of the high dimension input sequence. A part of the relevant features is represented by the magnitudes of the sequence elements. We expect waiting time to be higher when most service times are high and most interarrival times are low. Also, waiting times for each priority would be higher, the more customers of the highest priority appear in the sequence. Overall mean waiting time, however, in a "head of the line" discipline would remain unaffected—another analogy to the professor who switched from one university to another and raised the average level in both universities! A natural representation of these attributes would be by their means: $\overline{IRT}$—mean interarrival time, $\overline{SRV}$—mean service time, and $N_1$—observed number of first priority customers. Unfortunately, but expectedly, these are all very poor control variates, accounting for only a

few percentage points of the estimators variance. The reason is that the variates do not represent at all the permutations in the sequence, which are the main source of variation.

In trying to account for the permutations, a possible attempt would be to match service times and interarrival times in pairs.

For example, consider the recursive relation for single server systems

$$W_j = \max (W_{j-1} + SRV_{j-1} - IRT_{j,j-1}, 0)$$

We can make a try by substituting $W_{j-1}$ with a constant, and defining

$$V (\text{Cust } j) = \max (SRV_{j-1} - IRV_{j,j-1}, R) \quad .$$

Where $R \leq 0$ is the best estimation possible, and when $P [W = 0]$ is substantial, $R = 0$ is a very convenient choice. Now we define

$$V_t = \frac{1}{m} \sum_{\text{Cust } j,\ \text{Set}_t} V (\text{Cust } j) \quad ,$$

and

$$V_{tp} = \frac{1}{N_{tp}} \Sigma V (\text{Cust } j), \quad \text{where} \quad p = 1, 2 \qquad (22)$$

This would not be effective for the lower priority classes.

$\overline{V}_1$, $\overline{V}_2$ would be the control variates for the mean waiting times of the corresponding priority classes. $\overline{V}$ can serve not only for the overall mean waiting time but also for all of the priority classes.

The distribution of the $V$'s is easy to find; as the mean of independent and identically distributed random variables, the distribution of each is the truncated convolution of the distributions of service and interarrival times.

For example, if arrivals are Poisson and service time is exponentially and identically distributed, we get

$$F_{v+}(s) = 1 - \frac{\rho^2}{1+\rho} e^{-\mu s} \qquad s \geq 0$$

$$F_{v-}(s) = \frac{1}{1+\rho} e^{\lambda s} \qquad R \leq s \leq 0$$

$$= 0 \qquad s < R \quad ,$$

which yield

$$E[V] = \frac{1}{\mu\rho(1+\rho)} [\rho^2 - (1 - e^{\lambda R})] \quad ,$$

$$\text{Var}[V] = \frac{1}{\mu^2\rho^2(1+\rho)^2} \{2(1+\rho)[(1+\rho^3) - (1+\lambda R)e^{\lambda R}] - [\rho^2 - (1 - e^{\lambda R})]^2\} \quad ,$$

and for $R = 0$

$$E[V|R = 0] = \frac{\rho}{\mu(1+\rho)}$$

$$\text{Var}[V|R = 0] = \frac{\rho(2+\rho)}{\mu^2(1+\rho)^2} \quad . \tag{23}$$

The corresponding first moments for $V_t$ are immediate (because of the independence) and for practical values of $mV_t$ can be considered normally distributed (this assumption is not actually necessary for the use of $V_t$ as a control variate, as only the first moments are used). For example:

$$E[V_t] = E[V(\text{Cust})]$$

$$\text{Var}[V_t] = \frac{1}{m} \text{Var}[V(\text{Cust})] \quad .$$

Generally speaking, these moments do not exist for $V_{tp}$, as $P[N_{tp} = 0] > 0$. They do exist, however, for the conditional random variable $[V_{tp}|N_{tp} > 0]$. Clearly,

$$E[V_{tp}|N_{tp} > 0] = E[V(\text{Cust})] \quad .$$

The derivation of the variance is not computationally easy. But for Poisson arrivals, the normal approximation with large enough $m$ can be used together with a series expansion to get

$$\text{Var}\ [V_{tp} | N_{tp} > 0] \ \tilde{=}\ \frac{\text{Var}\ [V(\text{Cust})]}{E\ [N_{tp}]} \left\{ 1 + \frac{P_r[\text{Cust} \in C_p]}{E\ [N_{tp}]} + \ldots \right\} \quad .$$

Thus $V_{tp}$ can be used as a legitimate control variate.

A modification of $V_{t1}$ can be used to obtain indication if priority 1 customers arrived in groups, causing a substantial queue to be formed, rather than "evenly" distributed. That would be

$$V_{11}\ (\text{Cust}\ j) \ = \ \max\ (SRV_{j-1} - IRV_{j-1,j},\ 0)$$

where $j - 1$, $j$ are two consecutive arrivals of the priority class 1 (regardless of any other arrivals in between).

In searching for a good complement to "type $V$" control variates, note the low correlation between waiting time and either mean interarrival time or mean service time. Subsequently, there is practically no correlation whatsoever with

$$\overline{SRV} \ = \ \theta \overline{IRT}$$

where $\theta$ is a constant. Yet, the correlation of this compound variable with $V$ is not eliminated because the influence of $IRT$ in $V$ is truncated and it can be used as a complement of the type described above. There is no need to compute separately the optimal value for $\theta$; rather, the multiple control variates model can be used with $V$, $SRV$, and $IRV$, and all optimal coefficients are computed simultaneously. For example, in a simulation of a three priority, single server exponential system, the use of $V_t$ alone indicated a reduction of the variance to about 40% of its initial direct simulation value. But combined with $SRV$ and $IRT$ the indicated reduction was down to about 8% of the initial value. Although the indicated efficiencies in this case may be slightly higher than what could be regularly expected with these control variates (as would be explained below) the relation between the two contributions is significant.

A very convenient "marginal" control variate is $N_{t1}$. As noted, the higher $N_{t1}$, the higher would be the expected mean waiting times for each of the priority classes. Admittedly the correlation is low and would

usually account for only about 2% of the initial variance, but its importance stems from the fact $N_{t1}$ is in many cases independent of $V_t$ or $V_{tp}$ (although a somewhat elusive dependence exists with $V_{t11}$ and the efficiency is also maintained in the margin. Furthermore, $N_{t1}$ is very convenient computationally as a control variate of $X_1$. Practically all of the additional computation involved in using a control variate is in the estimation of the correlation with the estimated variable, the variance of the control variate and its covariance with other control variates being known. In the case of $N_{t1}$ and $X_t$, this does not involve even extra multiplications, as the product $X_{t1}N_{t1}$ is available even before $X_{t1}$ [see its Eq. (16)].

There is another control variate which is particular to the problem of priority queues. In many cases, the difficulty lies only with the priority discipline (not with the interarrival and service times stochastic mechanism). Analytical results are thus available for the overall mean waiting time (this does not hold for interrupt-repeat disciplines). There is a positive correlation between the observed mean waiting time for each of the priorities and observed overall mean waiting time. Therefore, $\overline{X}$ becomes a natural control variate.

It should be pointed out here that when $X$ was added to a set of control variates of the types described above, the increase in efficiency for the high priority classes was small. This means that little is gained by analyzing the overall process. It also indicated the effectiveness of $V$ as a control variate. $X$ is most effective with respect to a priority class, or a set of priority classes grouped together, that constitutes a large percentage of the total load. Also, its effect in the margin is obvious. For the lowest priority class, $X$ is extremely efficient, $V_{x,x_{NP}}$ often being in the order of 0.98.

A remarkable feature of control variates that has been observed, but not proved, is the fact that not only is the variance of the estimator reduced, but also the estimate of the variance is a much better one. Since

$$\hat{\sigma}^2_{w*} = \hat{\sigma}^2_x - \hat{CV}\,{}^t\Sigma^{-1}\hat{CV} \quad ,$$

this indicates that $\hat{\sigma}^2_x$ and $CV^t\Sigma^{-1}\hat{CV}$ are positively correlated random variables and that the correlation principle of the control variates operates

also in the estimate of the variance. As a good estimate of the confidence interval is an essential part of the experiment, this property is very significant.

The use of the control variates method may be extended to variables of which only the mean is known. In this case $\Sigma$ has to be estimated as well and the problem of bias is amplified. Here, too, it can be resolved by an estimation from a preliminary experiment, or extrapolation from a previous experiment with similar load. As long as the estimate of the optimal value for the vector of coefficients is independent, bias is not introduced. An error in $\alpha$ is thus never catastrophic, as at most, some efficiency will be lost. This loss is of the second order only, because $\sigma_w^2$ is convex in $\alpha$.

In conclusion, the control variates method was found quite effective for priority queues. The additional computation is small and just a few carefully chosen control variates will eliminate a substantial part of the initial variance. In a small number of exploratory trials with no more than four control variates the variance ratio was over 2 under "unfavorable" circumstances and over 10 under "favorable" ones.

*d.* STRATIFICATION

The basic idea is to divide the observation space into a set of mutually exclusive and jointly exhaustive strata, depending on one or more "stratification variables." We must estimate that

$$WS_k = E[X|X \in S_k] \qquad \text{for each stratum}$$

by

$$\bar{X}_k(X \in S_k), \qquad \text{or for any improved estimator}$$

denote

$$P_k = P[X \in S_k] \quad .$$

The estimator for $EX$ would be

$$W = \sum_k P_k WS_k$$

and

$$\text{Var } [W] = \sum_k P_k^2 \text{ Var } [WS_k] \quad .$$

Similar to what was done before (for $V_{tp}$), this is approximated by

$$\text{Var } [W] = \sum_k P_k \text{ Var } [X | X \in S_k] \quad .$$

Here $T$ (more important then $m$) has to be large enough.

We see that this is the mean "inter strata" variance, and that the "between strata" variance has been eliminated.

For a variable to be eligible as a stratification variable, not only the mean and variance have to be known in each of the strata, but also the distribution. With $m$ large enough this is not a serious restriction because for most of the set statistics normality can be assumed (unless autocorrelation is exceedingly high). The effectiveness of the method depends on finding a partition with the "between strata" variance a large portion of the total.

From the variables tested, none directly qualified in this respect. The typical distribution was as follows:

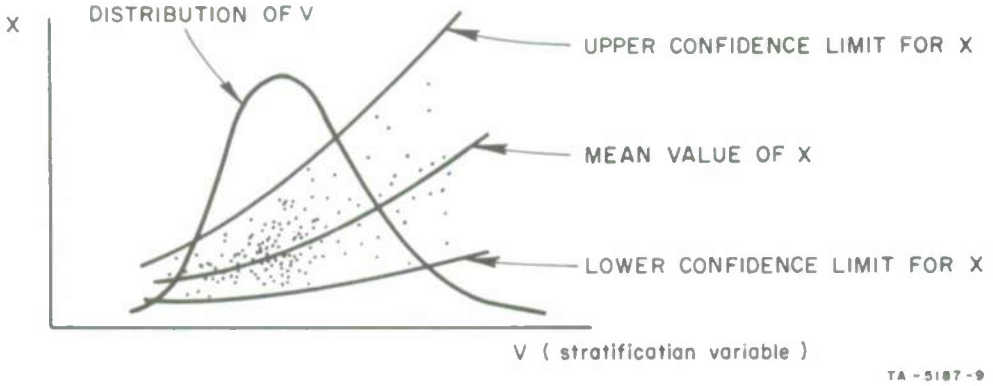



FIG. IV-1 TYPICAL DISTRIBUTION OF STRATIFICATION VARIABLE

We see that most of the variance actually lies in the strata with high $V$, and that the between strata variance is small. The situation is much worse for "bad" control variates (like $N_1$).

The method could still be useful if the number of observations in the important strata could be artificially increased by manipulating the input. For example, the simulation could be run with higher load than the one investigated, since the weights $P_k$ are analytically computed and do not depend on the number of observations. Unfortunately, this cannot be done under the present experimental setup where the end point of one set serves as an initial point for the next. The weights $P_k$ would compensate for bias in the input sequence, but not for the bias in the initial point.

Although some reduction in variance was obtained, on the whole the method was not found too efficient for priority queues. Not only the mean but also the variance have to be estimated in each of the strata. The reduction was usually not more than just a few percent (that could not be kept on the margin). It would be more effective in the simulation for transient results, where the initial point has to be considered specifically anyhow.

There is one exception. In many situations the probability of having to wait is known. We can estimate $E\ [X|X > 0]$ by $\overline{X}_+$, the average over positive values only, and then estimate mean waiting time for each priority class by

$$W = P[X > 0]\overline{X}_+$$

which has a smaller variance than $\overline{X}$. For a "head-of-the-line" priority queue discipline in an exponential system, the theoretical variance ratio would be approximately

$$\psi \cong \frac{1}{2 - \rho}$$

which was also observed in actual experiments.

This is not a very high reduction, especially for the more frequently encountered values of $\rho$. But here again there is practically no extra work—one comparison to zero is added for each customer, but the number of additions is reduced. A very slight increase is caused by the introduction of a counting index, altogether no change was observed in computing times.

The main question is whether this reduction would be maintained (proportionally) on the margin. In other words, would the use of control variates be as effective with respect to $X_+$ as it is for $X$? If so, one can obviously replace the estimator by

$$W = P[X > 0]\overline{W}_+$$

where $\overline{W}_+$ is some improved estimator (though control variates) of $E[X|X > 0]$. It seems it would be better to use a condition on the control variate too in this case, namely $V_+$ = average of $V$ only over values such that $V(\text{Cust}) > R$. The few experiments we held indicated the control variates method works just as well with $X_+$ as it does with $X$ for the exponential systems, "head-of-the-line" discipline over a wide range of $\rho$. It may or may not hold for other systems and other disciplines.

A different stratification operator, also involving manipulation of the input, will be discussed in connection with the antithetic variates.

*d.* ANTITHETIC VARIATES

The basic idea is this. $E[X]$ is to be estimated, and $Y$ is a random variable such that $E[Y] = E[X]$. Then

$$W = \frac{1}{2}(X + Y)$$

is an unbiased estimate of $E[X]$.

$$\text{Var } W = \frac{1}{4}\text{ Var } X + \frac{1}{4}\text{ Var } Y + \frac{1}{2}\text{ Cov } [X,Y]$$

When Var $X$ = Var $Y$

$$\text{Var } W = \frac{1}{2}\{\text{Var } X + \text{Cov } [X,Y]\} = \frac{1}{2}\text{ Var } X(1 + V_{xy}) \quad .$$

Since we have doubled the number of observations (observing also $Y$), this pays if $V_{xy}$ is negative. The $V_{xy}$ is the efficiency fraction.

As pointed out in Hammersley and Handcomb,[5] the antithetic variate method is awkward when applied to problems of high dimensionality, as is the case with priority queuing systems. One possible approach is to apply linear transformations to each of the individual terms in the input sequence. Page did this for a no-priority queuing system,[7] using either the transformation

$$\xi' = 1 - \xi \qquad \eta' = 1 - \eta$$

or the transformation

$$\xi' = \eta \qquad \eta' = \xi$$

or both. $\xi$ denotes the random numbers that determine interarrival time and $\eta$ the numbers that determine service times. The efficiency of the transformation stems from the fact that waiting time is monotone increasing with service time, and monotone decreases with interarrival time. In turn, service time and interarrival time are both monotone with the random numbers that generate them. The reported efficiency ratios were in the order of one-half. In a similar experiment we held for a priority queuing system, the observed efficiencies were of the same order for the overall mean waiting time. The gain was considerably smaller, however, for the individual priority classes. The reason may be that the different runs are antithetic only with respect to their "magnitude" of the input sequence, but not necessarily with respect to the permutations, which account for most of the variance. In the priority case, the permutations have an even stronger impact on the results for the individual classes, hence, the smaller gain.

A further transformation is presented in Hammersley and Handcomb.[5] This consists basically of stratification of the unit interval. The original suggestion was $\xi' = (\xi + k)/K \qquad k = 0, 1, \ldots, K-1$, but this is not directly applicable to the queuing problem, where in any given sequence each of the values has to be uniformly distributed in (0,1). The modified transformation $\xi' = (\xi + k/K) \bmod 1$, $k = 0, 1, \ldots, K-1$ is legitimate. This is particularly useful when the function is symmetric in the unit interval, which can be achieved by taking

$$f' = \frac{1}{2}\,[f(\xi') + f(1 - \xi')] \quad .$$

The natural way would be to do this separately for $\xi$ and $\eta$. For each value of the pair $(k_\xi, k_\eta)$ we may have four runs:

(1) $(\xi', \eta')$;

(2) $(1 - \xi', \eta')$;

(3) $(\xi', 1 - \eta')$;

(4) $(1 - \xi, 1 - \eta')$ .

This amounts to $4K^2$ runs.

Here a problem of scale exists. The efficiency of this transformation theoretically rises strongly with $K$. But we have $4K^2$ runs, the length of each is bounded from below. Therefore, to be very efficient, the experiment has to be very large, possibly larger than would be necessary by the requirements on the results. For this reason, really large experiments were not attempted.

An alternative approach rises from the nature of pseudo-random numbers used in computer work. All computer random number sequences are actually deterministic, entirely dependent on the initial choice. Observed mean waiting time for each priority class is therefore a positive, real, valued, function, albeit a complicated one, of a real number in (0,1)—the initial random number.

We can thus consider the problem as a problem of integration on (0,1) and avoid the high dimensionality altogether. The antithetic transformation alone has no value, because we now don't have monotonicity but the $2K$ transformations as described above are perfectly adequate. It is not obvious that the function is continuous, but the congruence method of random number generation suggests this would be the case. This alternative approach was not tested experimentally, and might provide interesting basis for further study.

An important question is whether the use of antithetic (or otherwise dependent) runs precludes the use of other variance reduction methods, particularly control variates. One problem here is that the covariance matrix for the control variates ($\Sigma$) is not easy to develop analytically for averages over dependent runs. It may sometimes be useful to choose as control variates such functions of $\xi$ and $\eta$ that are invariant under the transformations used, even at the cost of some efficiency. For the four antithetic runs described above (without the stratification

operator), $|1/2 - \xi|$ and particularly $|1/2 - \eta|$ proved to be reasonable approximations. This approach was not investigated further. Indeed, one of its shortcomings is that it would involve substantial preliminary experimentation to get the proper insight.

A more direct approach is to estimate $\Sigma$ as well as $CV$ (the covariance vector). Here the question of bias becomes more acute, but it can again be resolved by a preliminary experiment. Even then, $\alpha\ \hat{\Sigma}^{-1}\hat{CV}$ is a biased estimate of the optimal coefficients, but again it involves at most some reduction in efficiency.

On the experiments carried out, the averages of the control variates over antithetic runs did maintain the correlation with the respective averages of mean waiting times. It seems, therefore, that both methods can be operated together, although not all of the individual efficiency ratios would be maintained.

The antithetic variates method (with or without the stratification operator) has a certain appeal as an apparently general purpose method, applicable to all kinds of systems and loads without having to bother about insight into the stochastic nature of the process. This is not always true. Even if individual terms have monotone relationship with the estimated variables, the structure for the highly dimensional input sequence as a whole is far more complicated, as indicated above. It may, under special circumstances, even happen that the runs are positively correlated, which would actually increase the variance.

### e. SIMULTANEOUS ESTIMATION FOR ALL CLASSES

From a typical experiment, estimates $W_p$ for each of the priority classes as well as $W$ for the overall are available. To be consistent, these estimates must satisfy

$$W \sum_{p=1}^{NP} \lambda_p = \sum_{p=1}^{NP} \lambda_p W_p \quad ,$$

which means that actually more estimates are available than estimands. This suggests the use of a standard regression model

let $H$ be a $NP \times (NP + 1)$ matrix

$$H = \begin{bmatrix} I \\ a \end{bmatrix}$$

where

$$a_j = \frac{\lambda_j}{\sum_{j=1}^{NP} \lambda_i} \quad , \qquad j = 1, \ldots, NP$$

let $E$ be $NP \times 1$ vector $E_p = E[W_p] = E[X_p]$

let $\tilde{W}$ be the $(NP + 1)$ vector of estimators,

$$\tilde{W}_p = W_p \quad , \qquad \tilde{W}_{NP+1} = W$$

and let $V\tilde{W}$ be the covariance matrix of the estimators

$$V\tilde{W}_{ij} = \text{Cov}\ [\tilde{W}_i, \tilde{W}_j] \quad ,$$

$$i, j = 1, 2, \ldots, NP, NP + 1 \quad .$$

Then

$$E[\tilde{W}] = HE \quad .$$

Therefore

$$\tilde{W}^* = \{H^T[VW]^{-1}H\}^{-1}H^T[V\tilde{W}]^{-1}\tilde{W}$$

is the best estimator of $E$ given $\tilde{W}$ and

$$\{H^T[V\tilde{W}]^{-1}H\}^{-1}$$

is the covariance matrix for the best estimators.

The observed reduction in variance in this step was small, often negligible for the higher priority classes. The additional work involves estimating the elements Cov $[X_{pi}, X_{pj}]$ and Cov $[X, X_{pi}]$, $i, j = 1, \ldots, NP$,

an operation the size of which depends on $T$ and $NP$, and inverting which depends only on $NP$. The method may be more useful, therefore, when the required accuracy (and $m$) is high.

It is often practical to regard the estimates for the higher classes as fixed, and employ the regression model only for $W_{NP}$ and

$$\left[\lambda W - \sum_{i=1}^{NP-1} \lambda_i W_i\right] .$$

This does not reduce the work in estimation, but the inverted matrix is $2 \times 2$. As the initial estimates for the high priority classes are far better, this approach yields practically the same efficiency as employing the larger matrix.

If the regression model is discarded, there is a choice for estimating mean waiting time for the lowest priority with $W_{NP}$ or

$$\left[\frac{\lambda}{\lambda_{NP}} W - \sum_{i=1}^{NP-1} \frac{\lambda_i}{\lambda_{NP}} W_i\right]$$

(or any convex combination of these—the optimal one would have been determined by the regression model). The decision can obviously be made after the results from the experiment are obtained. Usually, when $X$ is not one of the control variates the variance of $W_{NP}$ is much higher than that of the alternative estimator. When overall mean waiting time is known, the variance of the two estimators are of the same order, but $W_{NP}$ is still the worse. Also, the estimate of Var $[W_{NP}]$ should be treated with suspicion in this case, as small errors in the estimate of $r_{x,x_{NP}}$ result in substantial errors in the estimate of Var $[W_{NP}]$.

## 3. Summary

The object of this study was to investigate efficient methods for the simulation of queuing systems with priority disciplines to estimate mean waiting times under "steady state" conditions. The problem of estimating transient behavior was kept in the background. It seems, though, that most of the methods presented would be applicable for the transient case with but little modification—some of them may indeed prove even more efficient.

Part of the experiments were made on the IBM 7090 computer at Stanford University, and part on the Burroughs B5500 at SRI. For the 7090, the programming language was SIMSCRIPT, and for the B5500 an ALGOL simulation program already available at SRI was used. Due to budget limitations, only a few experiments were run for each method, and the results indicate orders of magnitude rather than precise conclusions. As the exact realized efficiency depends on the system to be simulated, the load and the specified accuracy requirements, this seems the only way for a general investigation of the nature of the present study.

The best results were obtained with the control variates method, even when no more than four control variates were used for each estimand. Conditioning on $X > 0$ provided a further reduction in variance at no extra cost. The effectiveness of the other methods seems to depend more on the size of the experiment and the required accuracy. The stratification method provides the weights for "importance sampling" manipulation of the input. That, too, would be more applicable when required accuracy is high, or for the transient problem, where initial position has to be considered anyhow.

The present study was restricted to a single service center (though not necessarily a single server). A natural extension would be to queues in a network. Obviously, most of the methods would have to be modified, and in particular new control variates and stratification variables would have to be found. As the stochastic mechanism is more complicated, this would probably be more difficult for the network problem, and it may be expected that the control variates method particularly would lose some of its relative advantage over the other methods. On the other hand, the idea of treating the problem as a function of the initial random number only gains some appeal.

# V GUIDES AND PROCEDURES FOR THE APPLICATION OF QUEUING MODELS

## A. INTRODUCTION

Large scale military information systems have a number of common characteristics. In these systems there are usually one or more interconnected information processing centers. Each information processing center is usually equipped with one or more data processing units (computers) and associated memory systems. The memory system usually contains a hierarchy of memories, such as the high-speed core or thin-film memory, the mass core memory, the drum memory, the disc memory and the tape memory. A typical hardware organization of the information processing center is as shown in Fig. V-1. Incoming messages arrive at

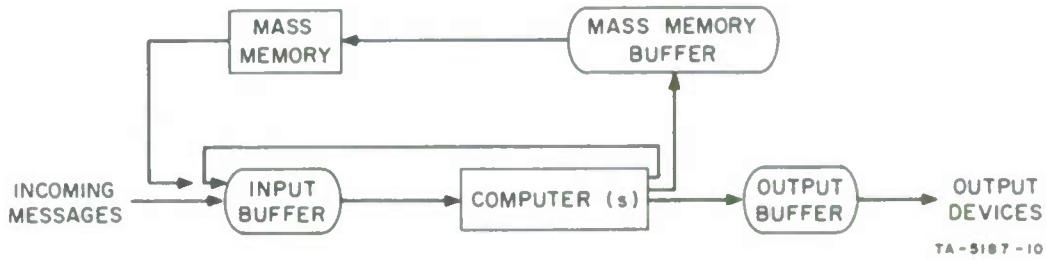



FIG. V-1 TYPICAL INFORMATION PROCESSING CENTER

the input buffer where they wait for the service of the computer. The incoming messages are usually categorized into classes, and the queuing discipline at the input buffer is usually of the priority type. An incoming message may require one or more consecutive services from the computer. The result of each computer service can be either a new request for computer service, or a request to retrieve information from the mass memory or an output message. An information retrieval request always waits in the mass memory buffer. The queuing discipline at the mass memory buffer can be the priority type. The result of an information retrieval is always a request for the computer service. The mass memory may contain several nonhomogenous memory units. Access to these memory units

can be made in parallel. The output messages queue up at the output buffer for transmission to the output devices.

The designer of such an information system is usually faced with the problem of determining the size and configuration of the buffers, the speed and the number of computers, the service time of the mass memory and the queuing or scheduling discipline required when the system parameters are not known exactly. As a matter of fact, during the early stages of system development the designer is often interested in investigating the expected performance and the probable weaknesses when the parameters vary over a given range. For example, during the early stages of system development the computer softwares are not fully developed, hence the computer service time characteristics are not precisely known. It therefore is quite desirable to establish the system response time as a function of computer service time. The designer might also be interested in determining the size of the buffers as a function of the size of softwares and the arrival rate of messages. To study the functional relationships between the system parameters and the system performance under a specific system hardware and software configuration the system designer usually establishes a model of the system and studies the functional relationships by manipulating the model.

The 473-L Simulation Model is a particular case of the generalized information system. The procedure used in analyzing the 473-L Simulation Model may therefore be used for analyzing the generalized information system. This procedure is described in the following section.

Although the variance reducing techniques, the control variates and the antithetic variates, have not been applied to the 473-L Simulation Model, the possible application of these techniques is indicated.

## B. GENERAL PROCEDURES

The prerequisite to a successful analysis of any information system is the establishment of a clear and precise description of the system from the point of view of servicing the incoming messages. The description should include the following items for each service station within a processing center:

(1) Input Message Characteristics.

(a) The classification of messages.

(b) The percentage mix of each class of messages.

(c) The mean and distribution of the inter-arrival intervals of each class of messages.

(d) The service requirement of each class of messages.

(e) The mean and distribution of the message lengths of each class of messages.

(2) Buffer (Queue) Characteristics.

(a) The organization of queues, *i.e.*, is the buffer divided into sections one for each class of message?

(b) The size of queue.

(c) The queuing (service) discipline.

(3) Server Characteristics.

(a) The number and type of servers.

(b) The mean and distribution of service times for each type of service.

(4) Output Characteristics.

(a) The destination of the output of a server.

The service stations in the typical processing center are the computer and the mass memory.

Having characterized the service centers, the operational measures to be obtained from the analysis must be defined. Some of the more commonly used operational measures are: the response time, the queue length and the utilization of server. We usually like to obtain the distribution of these measures, but in practice this is either too difficult or impossible to obtain. Hence, we are forced to settle for the mean or an extreme point of the distribution.

Having characterized the service centers and defined the operational measures desired, the next step is to construct a model of the processing center. The analytical modeling technique is always preferred over the numerical modeling (simulation) technique. Since the former technique can usually yield the analytical relationship between the system parameters and the system operational measures, the analytical relationship can then

be used in system evaluations such as cost-effectiveness studies. However, the analytical modeling technique is usually limited to the modeling of the simple service systems. Furthermore, the results of the analytical technique of modeling are usually limited to the first few moments of the operational measures. Nevertheless, the existence of even the first moment, the mean of the operational measure, may enable the information system designer to gain considerable understanding of the expected performance of the system.

The available analytical results of priority-queuing are listed in Reference 2. As far as it is known no significant new results have been added since the publication of this list. Any priority-queuing service center whose input, queue, and service characteristics do not meet those as listed cannot be analytically modelled. However, there are operational measures that do not require a knowledge of queuing discipline even though the service-center has a priority queuing discipline. For example, in deriving the mean utilization and the total mean storage of a service center it is not necessary to stratify the input into priority classes. Thus the nonpriority analytical results are applicable. If for some reason the storage is compartmentalized, one compartment for each priority class, then it is necessary to derive the mean storage for each priority class. In such a case, the priority-queue results must be used.

A typical information processing center usually contains more than one service center. Incoming messages usually require a sequence of services from these service centers. Each service center usually contains a buffer where messages may queue up. If the buffer is small with respect to the expected loading of the service center, blocking can occur. Blocking is defined as the inability of a service center to service further messages because the buffer of the subsequent service center is full. The analytical investigation of such a service system is quite formidable if not impossible. However, in practice the buffer is usually quite large and for all intents and purposes the buffer can be assumed to be infinitely large. It is thus possible to analyze each service center independently.

Frequently the information system is too complex for analytical modeling. Thus, we must resort to the simulation modeling technique. As was pointed by Conway,[4] there are three phases in an investigation by simulation that take place after the problem has been identified and a model formulated:

(1) Model implementation—description in a language acceptable to the appropriate computer.

(2) Strategic planning—design of experiment that will yield the desired information.

(3) Tactical planning—determination of how each of the test runs specified in the experimental design is to be executed.

There exists a number of simulation-oriented computing languages, among them are the SIMSCRIPT, the General Purpose System Simulation (GPSSII), the Control and Simulation Language (CSL) and the MILITRAN. No attempt has been made to evaluate the suitability of these languages to the simulation of information systems. In this project the original simulation model was programmed in ALGOL, since it was felt the model under study was relatively simple and would be more efficient to use ALGOL. During the strategic planning and tactical planning phases, however, it was necessary to make frequent changes in the sampling schemes. An ALGOL simulation program is not the easiest program for extensive modifications. Thus, the simulation model was reprogrammed in SIMSCRIPT during the later part of this project. The reason for choosing SIMSCRIPT over the other simulation languages are:

(1) SIMSCRIPT is the most commonly used simulation language, and

(2) a SIMSCRIPT compiler is available with the Stanford's 7090 computer.

An aspect of the experimental design for simulation experiment is reduction of the variance of a fixed volume of sampling without corresponding increase in computation labor. Three variance reducing techniques have been discussed in Sec. IV-D, namely, stratification, control variates and antithetic variates.

The stratification technique requires a knowledge of the distribution of the stratification variable. Usually the distribution is not known. Although it may be possible to assume a normal distribution, it is not possible to estimate the confidence attributable to the results. This technique is thus deemed impractical for information system simulation.

There does not seem to be any set procedure in the application of the control variates technique to the simulation of information systems. It usually involves a careful examination of the probabilistics structure of the simulated process. The purpose of the examination is to determine the

specific input or the combination of inputs—say $V$—in the simulation that clearly contributes to the variation in the output. Having found $V$, and knowing the distribution of the input(s) that make up the $V$, it is possible to derive the $E[V]$. Thus, the estimate of the output is $W = X - \alpha(V - E[V])$ where

$X$ is the simulated estimate of the output

$V$ is the simulated estimate of the control variate $V$

$E[V]$ is the theoretical mean of the control variate $V$

$\alpha$ is a constant.

A good control variate for a single exponential server priority queuing system has been found to be (Sec. IV-D):

$$V_{tp} = \frac{1}{N_{tp}} \sum_{\text{Cust } j \in \text{Set}_{tp}} V(\text{Cust } j) \quad .$$

where

$V(\text{Cust } j)$ = $\max (SER_{j-1} - IRV_{j,j-1}, R)$

$SER_{j-1}$ = the service time of the $j-1$-th customer

$IRB_{j,j-1}$ = the time interval between the $(j-1)$-th arrival and the $j$th arrival

$R$ = constant .

Other simple control variates are: interarrival time, service time, number of highest priority customer and the overall mean waiting time. We have found it possible to use more than one control variate.

In a more complex service system, say multiple parallel servers, a good control variate may be rather difficult to find. Thus, we may have to use the simple control variates and be contented with lower reduction in variance. For example, in the case of the 473-L Simulation Model the distribution of

(1) the interarrival interval of requests,

(2) the service time at the computer,

(3) the service time at the disc, and

(4) the number of disc searches,

can be used as the control variates for the mean response time measurement.

While as the distributions of

(1) the message types,

(2) the request message length,

(3) the interarrival intervals of requests,

(4) the search type,

(5) the data length,

(6) the program length, and

(7) the display message length,

can be used as the control variates for the mean storage size measurement.

Our investigation of antithetic variates technique has not advanced far enough to allow one to suggest any specific procedure for the application of this technique to information system simulation. However, the procedure outlined in Tocher[9] may be used, if there are $k$ distributions $R_1, R_2, \ldots, R_k$ to be used as sources of antithetic variates. In the 473-L Simulation Model, $k = 4$ was used for the mean response time measurement and $k = 7$ for the mean storage size measurement. It has been recommended that $k$ should be restricted to 2 or 3. Thus, we must select those distributions that affect the measurement most directly. In the case of mean storage size measurement, the interarrival intervals of requests, the search type, and the program length distributions are the three distributions that are most directly correlated to the measurement. Therefore these three distributions should be used as the primary control variables. Choose a sample size $n \cdot 2^k$. In each of $n$ sets of $2^k$ factorial experiment. Associate a vector $V_1, V_2, \ldots, V_k$ with each sample. If $V_i = 0$, use the variable $\xi_i$ in forming $R_i$; if $V_i = 1$ use the variable $1 - \xi_i$ in the transformation to $R_i$.

For $k = 3$, the table which follows illustrates the arrangement.

Table VI

A DESIGN OF AN EXPERIMENT
FOR ANTITHETIC VARIATES SIMULATION

| DISTRIBUTION | VARIABLE | SAMPLE | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| $R_1$ | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| $R_2$ | 2 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 |
| $R_3$ | 3 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |

Associated with $R_1$ there are four antithetic pairs, namely, (1,2), (3,4), (5,6), and (7,8). Let their averages be denoted 1′, 2′, 3′, and 4′.

The pairs (1′,2′) and (3′,4′) are antithetic pairs from distribution $R_2$ and their averages form an antithetic pair for $R_3$. Let $z_{jk}$ be the sample value of the $j$th sample of the $k$th set, where $j = 1, 2, \ldots, 8$ and $k = 1, 2, \ldots, .$ Let $Z_k = (1/8) \sum_j z_{jk}$. Then the estimate is $\bar{Z} = (1/n) \sum_k Z_k$. The estimated sample error is found from $[\sum_k (Z_k - \bar{Z})^2]^{\frac{1}{2}}$ and the expected variability of the response is derived from

$$(1/8) \sum_j [\sum_k (z_i - \bar{Z}_j)^2]^{\frac{1}{2}}$$

where

$$\bar{Z}_j = \frac{1}{n} \sum_k z_{jk} \quad .$$

For those who are interested in the design of simulation experiment and the tactical planning of simulation experiment, References 3 and 9 should be consulted. Tocher,[9] states the whole field of experimental design for simulation experiments is in its infancy and offers a fertile field for further research. Therefore, more efficient procedures will certainly be developed in both the application of variance reducing techniques and in the tactical planning of simulation experiments.

# VI CONCLUSIONS AND RECOMMENDATIONS

## A. CONCLUSIONS

Often it is possible to idealize the relatively complex network of service centers of an information system into independent service centers. By assuming these to be exponential service centers it is possible to apply the existing queuing theory to model the information system. Although the operational measures obtained from the analytical modeling may not be exact, nevertheless they provide an order of magnitude estimate of the system performance, and also a check on the Simulation Model is obtained.

Analytical modeling techniques are usually limited to the analysis of relatively simple queuing systems. Even then the analytical technique can, in most cases, yield only the mean of the measure. The distribution of the measure is usually left in the form of the transformation of the generating function. It would seem that a fruitful research effort is to deduce the approximate distribution from the transform of the generating function without actually carrying out the transform.

For more complex queuing systems we must rely upon the simulation techniques for the analysis of the system performances. The principal drawback of the simulation techniques are

(1) the amount of effort involved in the construction and check out of the simulation model, and

(2) the relatively large sample size required to obtain the operational measures that are sufficiently accurate.

With the appearance of the simulation-oriented computer languages, part of the first drawback has been overcome. The sample size required in the simulation can be drastically reduced by careful planning of the simulation strategies and tactics. The whole field of simulation strategies and tactics is still in its infancy and offers a fertile field for further research.

## B. RECOMMENDATIONS

It seems that the simulation modeling technique is going to be the principal technique for analyzing the queuing situations in information systems, unless there is a major breakthrough in the analytical technique which is not very likely. Therefore, it is recommended that future research effort in the application of queuing theory to information systems design be directed to the improvement of the simulation efficiency. In particular, the application of the control variates and the antithetic variates method in the simulation of a network of service centers should be investigated.

The derivation of the approximate distribution from the transformation of the generating function should be a worthwhile research effort, although it is not yet known how this can be accomplished.

*APPENDIX A*

# QUEUING ANALYSIS OF THE SYSTEM

## I INTRODUCTION

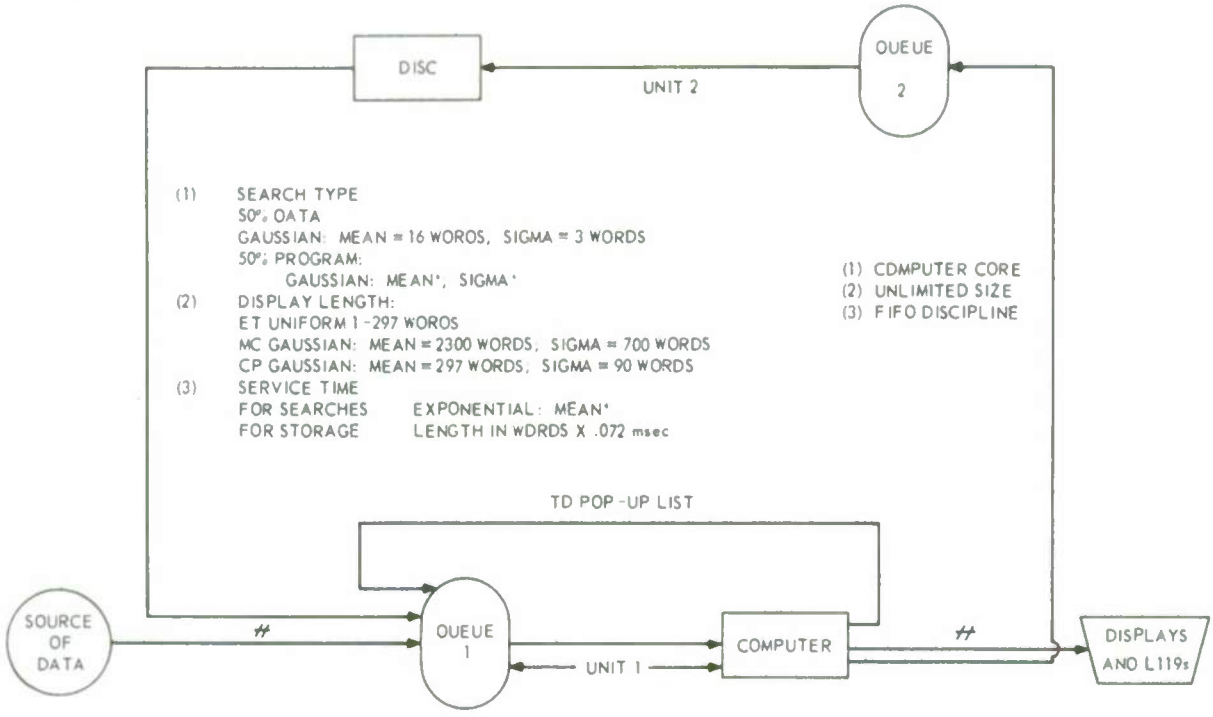
The principal purpose of the queuing analysis of the 473-L system is to investigate the possibility of applying the queuing theory to obtain some of the queuing measures of a real information processing system. The 473-L system model has been chosen for this purpose because a simulation model of this system has been constructed and simulated results of this system exist. Thus, it is possible to check the analytical results with the simulated results.

The 473-L system model is extracted from the monthly status report previously cited.[1] In this analysis, the analytical expressions for the utility of the computer, the utility of the disc, the mean response time, and the mean storage size were obtained.

## II DESCRIPTION OF 473-L SIMULATION MODEL

Figure A-1 shows the 473-L Simulation Model. Message requests from the "Source of Data" arrive at Buffer 1, $Q$-1, in random time—that is, the interarrival time is exponentially distributed with mean arrival rate of $\lambda$. For simplicity, message requests shall be called new requests, $R$. The new requests are categorized according to their priority level and message type. There are three priority levels and three message types; thus $R_{ij}$ is the $i$th priority level and $j$th message type request, where $i = 1, 2, 3$ and $j = 1, 2, 3$. In Reference 1, the $j = 1$ message type is called the electronic display type (ET), the $j = 2$ message type is called the multicolor display type (MC), and the $j = 3$ message type is called the console printout type (CP). The $i = 1$ request is the highest priority request and $i = 3$ is the lowest priority request. All arriving requests are stored in $Q$-1, which is a part of computer core.

New requests will always interrupt the computer operation. If the new request is of lower or equal priority to the one that the computer




SOURCE:

(1) TIME BETWEEN ARRIVALS
EXPONENTIAL: MEAN*
(2) THREE LEVEL PRIORITY
63.2% LOW
26.3% MEDIUM
10.5% HIGH
(3) DISPLAY OR MESSAGE TYPE
80% ET OR L119
10% MC
10% CP
(4) REQUEST LENGTH
50% 2 WORDS
50% UNIFORM 1-297 WORDS

BUFFER:

(1) COMPUTER CORE
(2) UNLIMITED SIZE
(3) PDP-UP LIST LENGTH
(a) FOR TASK IN PROCESS
# SEARCHES X 5 WORDS
(b) FOR REQUESTS IN QUEUE
# REQUESTS X 1 WORD
(4) PRIORITY ORDERING IN BUFFER

COMPUTER:

(1) SERVICE TIME
EXPONENTIAL: MEAN*
(2) SEARCHES
25% 0
75% GAUSSIAN: MEAN*
SIGMA*
(3) ALL BUFFER INPUTS CAUSE HOUSEKEEPING INTERRUPT TIME 1 msec

OUTPUT:

READOUT RATE
ET = 0
MC = 5.6 msec X # WORDS
CP = 560 msec X # WORDS

// RESPONSE TIME IS TIME DELAY FROM POINT A TO POINT B

* PARAMETERS VARIED
(1) MEAN TIME BETWEEN ARRIVALS: 2, 3, 4, 5, 6, 7, 9, 12 SEC
(2) COMPUTER SERVICE TIME: 40, 50, 65, 75, 80, 90, 180, 200, 220 msec
(3) SEARCHES: MEAN = 9, 15, SIGMA = 3, 5
(4) PROGRAM LENGTH: MEAN = 1000, 2500 WORDS, SIGMA = 300, 600
(5) DISC SERVICE TIME: 90, 180, 360, 540 msec

TM-5187-2

FIG. A-1 473-L SIMULATION MODEL

is working on when interrupted, the new request is routed to Buffer 2, $Q$-2, which is the other part of computer core. This new request is labelled as a pop-up request and will wait in $Q$-2 until it is read into the disc, under the FIFO (First in, First out) discipline. At the same time, a pop-up entry is added to the pop-up list in $Q$-1. If the new request is of higher priority to the interrupted message, the computer will analyze the request and initiate the first disc search, if any, required to fulfill the request. In all cases of interruption, the interrupted request is put back in $Q$-1, at the head of its priority queue.

Twenty-five percent of the requests require no disc search, they are the ET type requests. These requests are denoted by $R_{ij0}$. The remaining requests have a disc search distribution of the Gaussian type with mean $n$ and variance $\sigma_d$. These requests are denoted by $R_{ijn}$. The $R_{ij0}$ has higher priority than $R_{ijn}$ for $i = 1, 2, 3$. The $R_{ijn}$ requires one additional search to retrieve intermediate answers and format the reply message.

The new request whose first search has been initiated is placed at the end of its priority queue in $Q$-1, and this request is labelled as the "repeat" request. The computer will accept the first request in the highest priority queue in $Q$-1 as the next request to be analyzed. When the computer accepts a "repeat" request that requires further disc search, it will initiate the next disc search and place the request at the end of its priority queue in $Q$-1. The label of this request remains as the "repeat" request. If no further disc search is required, the computer will initiate the last search and label this request as the "computer" request and place the request at the end of its priority queue in $Q$-1. When the computer accepts a pop-up entry, the computer initiates a disc search to retrieve the pop-up request. Thus, a pop-up request requires one additional disc search.

The disc search requests and the pop-up store requests are queued up at $Q$-2 in the order of arrival. When a search is completed, the request, a "repeat" request, is routed to $Q$-1. If the "repeat" request has higher priority than that of the request being processed in the computer, the computer process is interrupted and the request in the computer is removed from the computer and is placed in $Q$-1. The processing of the "repeat" request is started. Otherwise, the "repeat" request is placed at the end of its priority queue in $Q$-1.

The "repeat" request returning from the disc is first analyzed by the computer. If it requires further search, then a disc search request is generated and placed in $Q$-2. If it requires no further search, this means a reply message is ready for transmission; the reply message is transmitted to the output devices, representing both displays and L119s.

The time when the message request arrived at $Q$-1 to the time when a reply message is ready for transmission is defined as the response time. If the reply message is the MC or CP type, it has additional transmission delay that is added to response time to give turn-around time.

## III SYSTEM PARAMETERS

### A. Input

As mentioned before, the interarrival time of new requests is assumed to be exponentially distributed with mean arrival rate of $\lambda$. The new requests are categorized according to their priority level and their message type. The percentage of requests that are the $i$th priority level is $Y_i$, where $Y_1 = 10.5\%$, $Y_2 = 26.3\%$, and $Y_3 = 63.2\%$. The percentage of requests that are the $j$th type is $W_j$, where $W_1 = 80\%$, $W_2 = 10\%$, and $W_3 = 10\%$. Fifty percent of the new requests have a length of two words; the other 50% have a length distributed uniformly between 1 and 297 words.

### B. Buffer ($Q$-1)

$Q$-1 is part of the computer core, the size of which we assume to be unlimited. $Q$-1 is used to store: (a) the new requests from the source of data, (b) the pop-up list, and (c) the results of disc search.

The pop-up list contains two types of request. The first type of request are those requests whose service has been initiated by the computer; the length of this type of message is the number of searches times five words. The second type of request are those requests that have been sent to the disc for temporary storage. Their length is one word. The second type of request becomes the first type of request when it has been retrieved from the disc for processing.

The length of the request returning from the disc search depends upon the type of disc search. If the search is for retrieving the requests in the pop-up list, the length distribution is the same as the new requests. If the search is for retrieving data, 50% of the searches are

of this type, the length is Gaussian distributed with mean $M_{ds}$ = 16 words and standard deviation $\sigma_{ds}$ = 3 words. The other 50% of the searches are the program retrieval type; their length is also Gaussian distributed with mean $M_{ps}$ and standard deviation $\sigma_{ps}$. Finally, the last search, the search to retrieve the intermediate answers and format the reply message, has a length distribution that depends upon the type of request: for the ET type the length is uniformly distributed from 1 word to 297 words; for the MC and CP types, the length has a Gaussian distribution of mean $\bar{l}_2$ = 2300 words and $\bar{l}_3$ = 297 words and standard deviation $\sigma_2$ = 700 words and $\sigma_3$ = 90 words, respectively.

All requests in $Q$-1 are arranged according to the priority level and in the order of the time of arrival at $Q$-1 within each priority level.

### C. Computer

The service time of the computer—the length of time required to analyze a request and initiate a disc search—is exponentially distributed with mean $h_c$. Twenty-five percent of the requests require no search; the other 75% have a Gaussian distribution of disc search with mean $n$ and standard deviation $\sigma_s$.

Each interruption of the computer operation consumes one msec of the computer time for the housekeeping function. The housekeeping time is symbolically represented by $h_h$.

### D. Buffer ($Q$-2)

The inputs to $Q$-2 are part of the output of the computer. The input consists of three types of request: the first is the storage type $R_s$, the second is the pop-up request retrieval type $R_r$, and the third is the disc information search type $R_d$. The message length of these three types of requests are as follows: $R_s$ has the same distribution as the new request to $Q$-1, $R_r$ is one word long, and $R_d$ is 5 words long.

The buffer size is assumed to be unlimited. Requests are arranged according to their order of arrival at $Q$-2.

### E. Disc

The disc is assumed to have unlimited size. The service time for storing a pop-up request $h_{ds}$ is proportional to the length of the request

or 0.072 msec × length. The service time for conducting a disc search is exponentially distributed with a mean $h_d$.

The output of the disc search is part of the input to $Q$-1.

## IV ANALYTICAL ANALYSIS

The 473-L Model is basically a tandem queue service system. The customers, new requests, may require successive services at the two servers, the computer and the disc. For example, a new request that requires $n$ disc searches must obtain service first at the computer and then at the disc, and then repeat this service sequence for $n + 1$ times before its service requirement is completed. Since the size of $Q$-1 and $Q$-2 has been assumed to be unlimited, the $Q$-1 and $Q$-2 are said to be statistically independent. A request with $n$ disc searches is said to have passed through $2(n + 1)$ statistically-independent stages of queue.

Throughout this analysis, we assume that the requests arrive at the $Q$-1 and $Q$-2 according to Poisson distribution. This assumption is justified by the following argument. The new requests arrive at the $Q$-1 according to Poisson distribution. The service time distribution of the computer is exponential. It has been shown in Burke,[12] that the interdeparture of a Poisson input and exponential service system is exponential. Thus, the interdeparture of requests from the computer is exponential. Since the inputs to the $Q$-2 are the outputs of the computer, the input at the $Q$-2 is Poisson. Now, the service time of the disc is assumed to be exponential;* thus, the interdeparture of requests from the disc is exponential. It follows that the return requests from the disc will arrive at the $Q$-1 according to Poisson distribution.

---

* The service time distribution of the disc is strictly speaking not Poisson. No more than 7% of the requests, depending upon the utility of the disc, are of the storage type. Fifty percent of this type of request requires a constant service time of 0.144 msec, and the other 50% has a uniform service time distribution from 0.072 msec to 21.38 msec. The other 93 or more percent of the requests has an exponential service time distribution with mean no less than 90 msec. Since the percentage of store request is relatively small, its influence on the overall service time distribution is negligible. Hence, the exponential assumption is a good first approximation.

The following notations are used in the analysis:

$\lambda$ = The mean arrival rate of new requests.

$R_{ijn}$ = The $i$th priority and $j$th type new request with $n$ disc searches, $i$ = 1, 2, 3 and $j$ = 1, 2, 3.

$Y_i$ = The percentage of new requests that are $i$th priority requests.

$W_j$ = The percentage of new requests that are $j$th type requests.

$M_{ds}$ = The mean of the Gaussian distribution of the retrieved length of the data type disc search.

$\sigma_{ds}$ = The standard deviation of the Gaussian distribution of the retrieved length of the data type disc search.

$M_{ps}$ = The mean of the Gaussian distribution of the retrieved length of the program type disc search.

$\sigma_{ps}$ = The standard deviation of the Gaussian distribution of the retrieved length of the program type disc search.

$\overline{l}_j$ = The mean length of the last disc search for the $j$th type requests.

$\sigma_j$ = The standard deviation of the length of the last disc search for the $j$th type requests.

$h_c$ = The mean service time of the computer.

$X$ = The percentage of new requests that require no disc search.

$n$ = The mean of the Gaussian distribution of the number of disc searches.

$\sigma_s$ = The standard deviation of the Gaussian distribution of the number of disc searches.

$h_h$ = The interrupt housekeeping time.

$R_s$ = The storage type requests at $Q$-2.

$R_r$ = The pop-up request retrieval type at $Q$-2.

$R_{ds}$ = The disc data search type request at $Q$-2.

$R_{ps}$ = The disc program search type request at $Q$-2.

$h_s$ = The mean service time for the $R_s$ requests.

$h_d$ = The mean service time for the $R_r$ or $R_{ds}$ or $R_{ps}$ requests.

$Y_i$ = The percentage of all requests entering $Q$-1 that are ith priority requests.

$\rho_c$ = The mean utility of the computer.

$\rho_d$ = The mean utility of the disc.

$N$ = The mean number of computer service units required by a new request.

$F_c$ = Frequency of computer service requests generated by a single new request.

$F_d$ = Frequency of disc service requests generated by a single new request.

$M_l$ = Mean length of the "new" request.

$M_r$ = Mean length of the repeat request.

$U$ = The probability that the first disc search is a data disc search.

$P_{di}$ = The probability that the ith disc search is a data type.

$P_{pi}$ = The probability that the ith disc search is a program type.

$V_n$ = The proportion of $n$ disc search requests that is the $R_{ds}$ type.

$M_{rn}$ = Mean length of $n$ consecutive "repeat" requests.

$M_{lc}$ = Mean length of a "computer" request.

$L_1$ = Total number of words of all requests generated by a weighted "new" request.

$L_{1w}$ = The weighted mean length per request at $Q$-1.

$L_{2w}$ = The weighted mean length per request at $Q$-2.

$L_2$ = Total number of words of all disc service requests generated by a new request.

$Z$ = The probability that a request, upon its arrival at $Q$-1, finds that the request that is being serviced by the computer has a lower or equal priority.

$T_r$ = The mean response time of the new requests.

### A. The Utility of the Computer

The mean utility of the computer, $\rho_c$, is defined as the percentage of time that the computer spent in servicing requests. The time that

the computer spent in servicing the requests consists of two parts: (a) the time spent in the analysis of requests, and (b) the housekeeping time when an interruption occurs. These two parts of computer utility are represented by $\rho_c'$ and $\rho_c''$, respectively.

The $R_{ij0}$ requests require only one unit of computer service for all values of $i$ and $j$. The $R_{ijn}$ request, for $n > 0$, requires $n + 2$ units of computer service if at the time the request arrives at $Q$-1 the computer is not busy, or if at the time the new request arrives at $Q$-1 the computer is busy and its priority is higher than that of the request that is being serviced by the computer. Thus the proportion of $R_{ijn}$ that requires $n + 2$ units of computer service is $(1 - \rho_c) + \rho_c Z = 1 - \rho_c(1 - Z)$. The rest of the $R_{ijn}$, $\rho_c(1 - Z)$ percent, requires $n + 3$ units of computer service.

The probability, $Z$, the sum of: (a) the probability that the new request is the $i = 1$ type and the request that is being serviced is either the $i = 2$ or 3 type, and (b) the probability that the new request $j = 2$ type and the request that is being serviced is the $j = 3$ type. Hence,

$$Z = Y_1(Y_2 + Y_3) + Y_2Y_3 \quad . \tag{A-1}$$

Since $X$ percent of the new requests are the $R_{ij0}$ type and $(1 - X)$ percent are the $R_{ijn}$ type, the mean number of computer service units, $N$, required by a new request is

$$\begin{aligned} N &= X + (1 - X)\{[1 - \rho_c(1 - Z)](n + 2) + \rho_c(1 - Z)(n + 3)\} \\ &= 1 + (1 - X)[n + 1 + \rho_c(1 - Z)] \quad . \end{aligned} \tag{A-2}$$

The mean service time for each unit of computer service is $h_c$, hence the mean service time for each new request is $h_cN$. For an arrival rate of $\lambda$, the utility of the computer for analyzing the request is

$$\begin{aligned} \rho_c' &= \lambda h_c N \\ &= \lambda h_c\{1 + (1 - X)[n + 1 + \rho_c(1 - Z)]\} \quad . \end{aligned} \tag{A-3}$$

Each new request will cause an interruption of the computer service if upon its arrival at $Q$-1 the computer is servicing another request. The probability that a new request finds the computer husy is $\rho_c$. For $\lambda$ new request arrival rate and $h_h$ mean housekeeping time per interruption, the time spent by the computer in housekeeping processing of new request is $\lambda h_h \rho_c$. Each new request will on the average generate $N - 1$ disc searches, where $N$ is given in Eq. (A-2). A disc search return request may interrupt the computer service if the priority of the return request is higher than that of the request that is being serviced. The probability that a return request finds the computer busy is again $\rho_c$, and the probahility that the return request is higher than that of the request that is being serviced is $Z$, as given in Eq. (A-1). Thus, for a new request arrival rate of $\lambda$, the return request interruption frequency is $\lambda\rho_c(N - 1)Z$. The time spent by the computer in housekeeping processing of the return request is then $\lambda h_h \rho_c (N - 1)Z$. Consequently, the total housekeeping load at the computer is

$$\rho_c'' = \lambda h_h \rho_c [1 + (N - 1)Z] \quad . \tag{A-4}$$

Combining Eqs. (A-3) and (A-4) one has

$$\begin{aligned} \rho_c = {} & \lambda h_c \{1 + (1 - X)[n + 1 + \rho_c(1 - Z)]\} \\ & + h_h \rho_c \{1 + (1 - X)[n + 1 + \rho_c(1 - Z)]Z\} \quad . \end{aligned} \tag{A-5}$$

By rearranging terms in Eq. (A-5), one obtains a quadratic equation in $\rho_c$, $A\rho_c^2 + B\rho_c + C = 0$, where

$$A = \lambda h_h (1 - X)(1 - Z)Z$$

$$B = \lambda\{h_c(1 - X)(1 - Z) + h_h[1 + (1 - X)(n + 1)Z] - 1\}$$

$$C = \lambda h_c [1 + (1 - X)(n + 1)] \quad .$$

Therefore,

$$\rho_c = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A} \quad . \tag{A-6}$$

However, for $A \ll B$ or $C$, which is the case under this investigation, Eq. (A-6) may be approximated by

$$\rho_c = -\frac{C}{B} \quad . \tag{A-7}$$

B. The Utility of the Disc

There are four types of requests that require the service of the disc; these are the store type requests $R_s$, the pop-up request retrieval type $R_r$, the disc data search type requests $R_{ds}$, and the disc program search type requests $R_{ps}$. As far as the utility of the disc is concerned, the $R_r$, and $R_{ds}$, and $R_{ps}$ requests may be treated as a single type, their service time being exponentially distributed with mean service time $h_d$. The mean service time for $R_s$ is $h_s$.

The $R_s$ is generated when the new request, which requires disc search, upon reaching $Q$-1 finds that the computer is busy, and that the priority of the new request is equal to or lower than that of the request in the computer. The percent of new requests that require disc search is $1-X$, the probability of finding the computer busy is $\rho_c$, and the probability of finding the request being serviced by the computer has a higher or equal priority is $(1 - Z)$. Therefore, the probability of a new request generates $R_s$ at $Q$-2 is $(1 - X)\rho_c(1 - Z)$. The disc utility due to $R_s$ per new request is $(1 - X)\rho_c(1 - Z)h_s$. The number of $R_r$, $R_{ds}$, and $R_{ps}$ generated by a single new request is $N - X$, where $N$ is given by Eq. (A-2). The disc utility due to the $R_r$, $R_{ds}$, and $R_{ps}$ per new request is $(N-X)h_d$. Therefore the total disc utility is, for a new request arrival rate of $\lambda$,

$$\rho_d = \lambda(1 - X)\{[n + 1 + \rho_c(1 - Z)]h_d + \rho_c(1 - Z)h_s\} \quad . \tag{A-8}$$

C. The Mean Response Time

The mean response time of a new request $T_r$ is defined as the mean elapsed time between the arrival of the new request at $Q$-1 to the time when it leaves the computer after the final disc search has been completed.

From the response time point of view, the new requests may be classified into three categories. The first is the requests that need no disc search $R_{i,j,0}$; $X$ percent of the requests are of this category. The

second is the $R'_{ijn}$ type of request that finds the computer not busy when it arrives at $Q$-1 or finds the computer busy but its priority is higher than that of the request in service. Thus, the percent of new requests that are in the second category is $(1 - X)[(1 - \rho_c) + \rho_c Z]$. The third is the $R''_{ijn}$ type of request that finds the computer busy; its priority is lower than or equal to that of the request in service. The proportion of new requests that belong to the third category is thus $(1 - X)\rho_c(1 - Z)$.

The mean response time of the first category of request $T_1$ is the mean waiting time $W_1$ plus the mean computer service time $h_c$ for the request. In this case, the mean waiting time is that of a single-exponential server system, namely

$$W_1 = \frac{\rho_c h_c}{1 - \rho_c} \quad .$$

Therefore,

$$T_1 = h_c \left[ 1 + \frac{\rho_c}{1 - \rho_c} \right]$$

$$= \frac{h_c}{1 - \rho_c} \quad . \tag{A-9}$$

The second category of requests require $n + 1$ successive services at the disc and the computer. Each service at the disc or the computer consists of a waiting time in $Q$-2 or $Q$-1 and a service time in the disc or the computer. The service time of the disc is exponentially distributed with mean service time $h_d$, and the disc is a single server. Thus, the mean delay experienced by a request at the disc is that of a single-exponential server, or $h_d/(1 - \rho_d)$. The mean delay experienced by the request at the computer is as given in Eq. (A-9). Since $Q$-1 and $Q$-2 are unlimited in size, the $Q$-1 and $Q$-2 are statistically independent of each other. Thus, the mean response time for the second category of requests $T_2$ is merely $n + 1$ times the sum of the mean delay at the disc and the mean delay at the computer, or

$$T_2 = (n + 1)\left[ \frac{h_c}{1 - \rho_c} + \frac{h_d}{1 - \rho_d} \right] \quad . \tag{A-10}$$

The third category of requests requires $n + 2$ successive services at the disc and the computer. In addition it requires one additional service at the disc, the storing of the request in the disc. The mean service time of this last type of disc service is $h_s$. Thus, the mean response time of the third category of requests $T_3$ is $n + 2$ times the sum of the mean delay at the disc, the mean delay at the computer plus one mean waiting at $Q$-2, and one disc service time whose mean is $h_s$ for storing the request, or

$$T_3 = (n + 2)\left[\frac{h_c}{1 - \rho_c} + \frac{h_d}{1 - \rho_d}\right] + \frac{h_s}{1 - \rho_d} \quad . \tag{A-11}$$

The total response time $T_r$ is the weighted sum of response time of each of the three categories, or

$$\begin{aligned} T_r &= XT_1 + (1 - X)[1 - \rho_c + \rho_c Z]T_2 + (1 - X)\rho_c(1 - Z)T_3 \\ &= \frac{Xh_c}{1 - \rho_c} + (1 - X)\left\{(n + 1)\left[\frac{h_c}{1 - \rho_c} + \frac{h_d}{1 - \rho_d}\right]\right. \\ &\quad \left. + \rho_c(1 - Z)\left[\frac{h_c}{1 - \rho_c} + \frac{h_d + h_s}{1 - \rho_d}\right]\right\} \quad . \end{aligned} \tag{A-12}$$

### D. The Mean Response Time of $i$th Priority Request

In this service system the incoming requests are categorized into three priority classes. All incoming requests require at least one initial service by the computer. $1 - X$ percent of all incoming requests require $n + 1$ subsequent services by the disc and the computer. Of the $1 - X$ percent requests, those requests that find the computer serving a higher or equal priority request when they arrive at the computer require an additional service by the disc and the computer. Since the probability of finding the computer busy is $\rho_c$ and the probability of an $i$ or higher priority request is being serviced is $\sum_{j=1}^{i} Y_j$, the probability of an $i$th priority request requiring an additional service is $(1 - X)\rho_c \sum_{j=1}^{i} Y_j$. The number of subsequent computer services required by an $i$th priority request is $(1 - X)(n + 1 + \rho_c \sum_{j=1}^{i-3} Y_j)$.

Let $R_i$, $i = 1, 2, 3$, be the first request for the service of the computer by the $i$th priority requests and $R_i$, $i = 4, 5, 6$, be the subsequent

request for the service of the disc and the compnter by the ith priority requests. Let $\rho_i$ be the loading of the compnter dne to the $R_i$. If $\lambda$ is the mean arrival rate of the incoming requests and $Y_i$ is the proportion of the requests that are of the ith priority, the loading dne to $R_i$ is

$$\rho_i = \lambda h_c Y_i \quad , \qquad i = 1, 2, 3 \quad , \tag{A-13}$$

$$\rho_i = \lambda h_c Y_i (1 - X)\left(n + 1 + \rho_c \sum_{j=1}^{i-3} Y_j\right) \quad , \qquad i = 4, 5, 6 \ . \tag{A-14}$$

In this analysis both the computer and the disc are assumed to be exponential servers. The queue discipline of the computer is the interrupt priority queue type, a $R_i$ can interrupt the service of $R_j$ if $i < j$. The queue discipline of the disc is the ordered queue type. The mean time spent by a $R_i$ request in waiting for and service by the computer is given as[2]

$$T_i' = \frac{h_c}{(1 - P_{i-1})(1 - P_i)} \tag{A-15}$$

where $h_c$ is the mean service time of the computer and $P_i = \sum_{k=1}^{i} \rho_i$.

The time spent by a request in waiting for and service by the disc is not distingnished by the priority type (the queuing discipline at the disc is the ordered queue type), but by the type of disc service required. From the service time point of view there are two basic types of disc service requests, the disc search type and the disc store type. The mean time spent by either of the two types of requests in waiting for the service by the disc is $(h_d\rho_d)/(1 - \rho_d)$, which is the mean waiting time of a single exponential server with mean loading of $\rho_d$ and mean service time of $h_d$. Let the mean service time of the disc search request and the disc store request be $h_{se}$ and $h_{sp}$ respectively. The mean time spent by a disc search request and a disc store request in waiting for and service by the disc are

$$T_d = h_{se} + \frac{h_d\rho_d}{1 - \rho_d} \tag{A-16}$$

and

$$T_p = h_{sp} + \frac{h_d \rho_d}{1 - \rho_d} \tag{A-17}$$

respectively.

An $i$th priority incoming request require an initial service by the computer, the mean response time of this service is $T'_i$. $1 - X$ percent of the $i$th priority incoming requests require $n + 1$ subsequent services by the computer and $n + 1$ disc search type service by the disc. The mean response time for the total of $n + 1$ services by the computer and the disc is $(n + 1)(T'_{i+3} + T_d)$. In addition $(1 - X)\rho_c \sum_{j=1}^{i} Y_j$ percent of the $i$th priority request requires a subsequent service by the computer and a disc store type service by the disc. Thus, the grand total response time of an $i$th priority request is

$$T_i = T'_i + (1 - X)[(n + 1)(T'_{i+3} + T_d) + \rho_c \sum_{j=1}^{i} Y_j (T'_{i+3} + T_p)] . \tag{A-18}$$

E. The Mean Storage Size

The mean storage size is defined as the average number of words stored in $Q$-1 and $Q$-2. Since the computer and the disc are the exponential type service system the mean number of requests in $Q$-1 is $L_{R1} = 1/1 - \rho_c$ and in $Q$-2 is $L_{R1} = 1/1\rho_d$. The mean number of words in $Q$-1 and in $Q$-2 are

$$L_{Q1} = L_{R1} \cdot \overline{L_1} = \overline{L_1}/(1 - \rho_c) \tag{A-19}$$

$$L_{Q2} = L_{R2} \cdot \overline{L_2} = \overline{L_2}/(1 - \rho_d) \tag{A-20}$$

respectively, where $\rho_c$ and $\rho_d$ are given by Eqs. (A-7) and (A-8). $\overline{L_1}$ and $\overline{L_2}$ are the mean length of a request in $Q$-1 and $Q$-2 respectively.

Before proceeding to find the $\overline{L_1}$ and $\overline{L_2}$, the relative frequency of occurrence and the mean length of various types of request are to be examined from the storage point of view. New requests can be classified into three categories, namely $R_{ij0}$, $R'_{ijn}$ and $R''_{ijn}$, and their relative frequency of occurrence are $X$, $(1 - X)[1 - \rho_c(1 - Z)]$ and $(1 - X)\rho_c(1 - Z)$ respectively. Each of these three "new" requests generate other types of requests as they are processed by the computer and the disc. The $R_{ij0}$ request generates a "new" request, $R_n$, at $Q$-1 only. The $R'_{ijn}$ request generates a $R_n$ at $Q$-1 and at the same time generates $n + 1$ disc search

requests, $R_s$, at $Q$-2. Each of the first $n$ $R_s$ reqnests in term generates a "repeat" request, $R_p$, for $Q$-1. The last $R_s$ request generates a "computer" request, $R_c$, for $Q$-1. The $R''_{ijn}$ reqnest generates a $R'_n$ at $Q$-1, and at the same time generates a disc store request, $R_d$, for $Q$-2 and a pop-up list request, $R_p$, for $Q$-1. When the $R_p$ is processed by the computer a pop-up list retrieve request, $R_{pr}$, is generated for $Q$-2, which results in the retrieval of the stored $R'_n$ from the disc and is placed into $Q$-1. The retrieved $R'_n$ is processed as the $R_n$ generated by the $R'_{ijn}$ request. Table A-1 is a summary of the number of each type of request generated by the $R_{ij0}$, $R'_{ijn}$ and $R''_{ijn}$ request.

Table A-I

NUMBER OF GENERATED REQUESTS

| ORIGINAL REQUEST | | GENERATED REQUEST TYPE | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Type | Relative Frequency of Occurrence | $Q$-1 | | | | | $Q$-2 | | |
| | | $R_n$ | $R'_n$ | $R_p$ | $R_r$ | $R_c$ | $R_d$ | $R_{pr}$ | $R_s$ |
| $R_{ij0}$ | $X$ | 1 | | | | | | | |
| $R'_{ijn}$ | $(1-X)[1-\rho_c(1-Z)]$ | 1 | | | $n$ | 1 | | | $n+1$ |
| $R''_{ijn}$ | $(1-X)\rho_c(1-Z)$ | 1 | 1 | 1 | $n$ | 1 | 1 | 1 | $n+1$ |

Let $F_n$, $F'_n$, $F_p$, $F_r$, $F_c$, $F_d$, $F_{pr}$, and $F_s$ be the number of the $R'_n$, $R_p$, $R_r$, $R_c$, $R_d$, $R_{pr}$, and $R_s$ requests generated by an original request respectively. From Table A-I it is seen that

$$F_n = X + (1-X)[1-\rho(1-Z)] + (1-X)\rho_c(1-Z) = 1 \quad ,$$

and similarly

$$F'_n = F_p = F_d = F_{pr} = (1-X)\rho_c(1-Z)$$

$$F_r = n(1-X)$$

$$F_c = 1-X$$

$$F_s = (n+1)(1-X) \quad . \tag{A-21}$$

Let $F_1$ and $F_2$ be the number of requests, regardless of type generated by an original request at $Q$-1 and $Q$-2 respectively, then

$$F_1 = F_n + F'_n + F_p + F_r + F_c = 1 + (1-X)[n+1+2\rho_c(1-Z)] \quad , \tag{A-22}$$

and

$$F_2 = F_d + F_{pr} + F_s = (1 - X)[n + 1 + 2\rho_c(1 - Z)] \quad . \qquad (A\text{-}23)$$

Let $M_n$, $M'_n$, $M_p$, $M_r$, $M_c$, $M_d$, and $M_s$ be the mean length of the $R_n$, $R'_n$, $R_p$, $R_r$, $R_c$, $R_d$, and $R_s$ requests respectively. The mean length of $R_{pr}$ is the same as that of $R_p$. The weighted mean length per request at $Q$-1 is

$$\overline{L_1} = \frac{(F_n M_n + F'_n M'_n + F_p M_p + F_r M_r + F_c M_c)}{F_1} \quad , \qquad (A\text{-}24)$$

and at $Q$-2 is

$$\overline{L_2} = (F_d M_d + F_{pr} M_p + F_s M_s) F_2 \quad . \qquad (A\text{-}25)$$

Only the $M_n$ and $M_p$ are the basic system parameters. The other mean length of requests are to be developed in term of the basic system parameters, namely, $M_n$, $M_p$, $W_j$, $\overline{l_j}$, $M_{ds}$, and $M_{ps}$.

The expressions for $M_d$ and $M'_n$ are to be developed first. When a $R''_{ijn}$ is processed by the computer the original request plus the disc search address for each of the $n + 1$ disc searches, plus the disc store address are forwarded to $Q$-2 as a $R_d$. The mean length of each disc address is $M_p$, therefore the mean length of $R_d$ is

$$M_d = M_n + (n + 2)M_p \quad . \qquad (A\text{-}26)$$

When the $R_d$ is processed by the disc the request minus the disc store address is stored into the disc. This request when retrieved by a $R_{pr}$ becomes a $R'_n$, thus the mean length of $R'_n$ is

$$M'_n = M_d - M_p = M_n + (n + 1)M_p \quad . \qquad (A\text{-}27)$$

$M_c$ is the mean length of the display message or $R_c$ retrieved by the last $R_s$ from the disc. There are three types of display messages, their proportion and mean length are $W_j$ and $\overline{l_j}$, $j = 1, 2, 3$, respectively. Hence, the mean length of $R_c$ is

$$M_c = \sum_{j=1}^{3} W_j \overline{l_j} \quad . \qquad (A\text{-}28)$$

To develop the expression for $M_r$ and $M_s$ it is necessary to review the process of handling the $R_s$ and $R_r$ requests by the disc and the computer. The first disc search request, $R_{s1}$, has a search length of $L_{s1} = (n + 1)M_p$. When $R_{s1}$ is processed by the disc a message is retrieved from the disc. This message and the $R_{s1}$ (its search length reduced by one $M_p$ unit) are placed into $Q$-1 to form the first "repeat" request, $R_{p1}$. Thus, the length of $R_{p1}$ is $L_{r1} = nM_p + M_s$ where $M_s$ is the mean length of the retrieved message. The $R_{s2}$ is processed by the computer and is placed into $Q$-2 to become $R_{s2}$. From request-length point of view $L_{s2} = L_{r1}$. When the $R_{s2}$ is processed by the disc the message retrieved, plus its search length reduced by one $M_p$ unit, are placed into $Q$-1 to become the $R_{p2}$. The length of $R_{p2}$ is then $L_{r2} = (n - 1)M_p + M_s$. The $R_{p2}$ is processed by the computer and is placed into $Q$-2 to become $R_{s3}$. This process is repeated $n$ times. In general, the length of $R_{pr}$ is $L_{rk} = (n + 1 - k)M_p + M_s$, and $L_{rk} = L_{s(k+1)}$ for $k \geq 1$. Thus, the mean length of a $R_p$ is

$$M_r = \frac{1}{n}\sum_{k=1}^{n} L_{rk} = \frac{1}{n}\sum_{k=1}^{n}[n + 1 - k)M_p + M_s] = \frac{n + 1}{2} M_p + M_s \quad \text{(A-29)}$$

and the mean length of a $R_s$ is

$$M_s = \frac{1}{n + 1}\sum_{k=1}^{n+1} L_{sk} = \left\{\frac{1}{n + 1} \frac{(n + 1)(n + 2)}{2} M_p + nM_s\right\} . \quad \text{(A-30)}$$

There are two types of $R_s$, the data disc search type, $R_{ds}$, and the program disc search type, $R_{ps}$. The mean message length retrieved by $R_{ds}$ is $M_{ds}$, and that retrieved by $R_{ps}$ is $M_{ps}$. Let $V_n$ be the proportion of $R_s$ that are the $R_{ds}$ types, thus $M_s = V_nM_{ds} + (1 - V_n)M_{ps}$. Substitute the expression for $M_s$ into Eq. (A-29) and (A-30) one has

$$M_r = \frac{n + 1}{2} M_p + V_nM_{ds} + (1 - V_n)M_{ps} , \quad \text{(A-31)}$$

and

$$M_s = \frac{1}{n + 1}\left\{\frac{(n + 1)(n + 2)}{2} M_p + n[V_nM_{ds} + (1 - V_n)M_{ps}]\right\} . \quad \text{(A-32)}$$

The expression for $V_n$ is a function of the rules that determine the type of disc search for the $k$th disc search. The rules are as follows:[1*]

(1) the probability of $R_{s1}$ being a $R_{ds}$ is $U$ and being a $R_{ps}$ is $1 - U$,

(2) if $R_{sk}$ is a $R_{ps}$ then $R_{s(k+1)}$ must be a $R_{ds}$, but if $R_{sk}$ is a $R_{ds}$ then $R_{s(k+1)}$ can be a $R_{ds}$ with a probability of $U$ and be a $R_{ps}$,

(3) the $R_{sn}$ is always $R_{ps}$.

Let $P_{dk}$ and $P_{pk}$ be the probability that the $k$th disc search is a $R_{ds}$ and $R_{ps}$ respectively. Since a disc search can only be a $R_{ds}$, or a $R_{ps}$, therefore $P_{dk} = 1 - P_{pk}$. According to Rule-1 $P_{d1} = U$ and $P_{p1} = 1 - U$. According to Rule-2 $P_{d(k+1)} + P_{dk}U = 1 - P_{dk}(1 - U)$. Since $P_{pk} = 1 - P_{dk}$, therefore

$$P_p(k + 1) = P_{dk}(1 - U) \quad . \tag{A-33}$$

Equation (A-27) is a recurrence equation, once $P_{d1}$ is given $P_{dk}$ and $P_{pk}$ can be uniquely determined. The proportion of the $n$ disc search that are the $R_{ds}$ type is

$$V_n = \frac{1}{n}\sum_{k=1}^{n-1} P_{dk} \quad . \tag{A-34}$$

It is to be noted that the summation is over $n - 1$ since by Rule 3 the $R_{sn}$ is always a $R_{ps}$.

Substituting Eqs. (A-21), (A-22), (A-23), (A-26), (A-27), (A-31), and (A-32) into Eqs. (A-24) and (A-25), and then substituting Eqs. (A-24) and (A-25) into Eqs. (A-19) and (A-20), we get

$$L_{Q1} = \frac{M_n + (1-X)\left\{n\left[\frac{n+1}{2}M_p + V_n M_{ds} + (1-V_n)M_{ps}\right] + M_c + [M_n + (n+2)M_p]\rho_c(1-Z)\right\}}{(1-\rho_c)\{1 + (1-X)[n + 1 + 2\rho_c(1-Z)]\}} \tag{A-35}$$

---

* These rules are in accordance with that stated in the program listing of the "473-L Simulation" Monthly Status Report. The main text of the above cited report simply stated that the probability of a $R_s$ being $R_{ps}$ is $U$. In order that the analysis conforms with the actual simulation model, the rules as stated in the program listing are used.

and

$$L_{Q2} = \frac{\frac{(n+1)(n+2)}{2} M_p + n[V_n M_{ds} + (1-V_n)M_{ps}] + [M_n + (n+3)M_p]\rho_c(1-Z)}{(1-\rho_d)[n+1+2\rho_c(1-Z)]}, \quad \text{(A-36)}$$

where $M_c$ and $V_n$ are given in Eqs. (A-28) and (A-34).

In the simulation program listing of the "473-L Simulation" Report 1, it was stated the mean length of $R_{s1}$, due to a $R''_{ijn}$ request only, is the same as the mean length of $R''_n$, namely, $L_{s1} = M_n + (n+1)M_p$. We note also this mean length is one $M_n$ unit longer than that used in deriving Eqs. (A-35) and (A-36). Furthermore, this one unit of $M_n$ is not dropped from subsequent $R_p$, $R_s$, and $R_c$ requests. For example, the mean length of $R_{pk}$ is ${}_3L_{rk} = (n+1-k)M_p + M_s + M_n$, and the mean length of $R_c$ is $M_c = \sum_{j=1} W_j \overline{l_j} + M_n$. Using the new storage rule, which shall be called Rule 2 to distinguish it from the Rule 1 used in deriving Eqs. (A-35) and (A-36), the expressions for the mean number of words in $Q$-1 and $Q$-2 are

$$L_{Q1} = \frac{M_n + (1-X)\left\{n\frac{n+1}{2}M_p + V_n M_{ds} + (1-V_n)M_{ps}] + M_c + (n+2)(M_n+M_p)\rho_c(1-Z)\right\}}{(1-\rho_c)\{1+(1-X)[n+1+2\rho_c(1-Z)]\}}, \quad \text{(A-37)}$$

and

$$L_{Q2} = \frac{\frac{(n+1)(n+2)}{2} M_p + n[V_n M_{ds} + (1-V_n)M_{ps}] + (M_n+M_p)(n+3)\rho_c(1-Z)}{(1-\rho_d)[n+1+2\rho_c(1-Z)]}. \quad \text{(A-38)}$$

It is believed Rule 2 is not logical; nevertheless, Eqs. (A-37) and (A-38) are presented here so that they may be used to check with the results of simulation.

*APPENDIX B*

# CONFIDENCE INTERVALS FROM A SIMPLE SAMPLE

## 1 INTRODUCTION

Assuming sample sizes from the simulation are large enough to make average output approximately normally distributed, symmetric confidence intervals are desirable measures of variation in the estimates. This section is primarily concerned with obtaining the confidence interval.

### A. Confidence Intervals

In determining confidence intervals, assumptions are:

(1) The transition probabilities are ergodic. This is necessary because only one simulation run will be made from a fixed initial state.

(2) The initial state can be chosen so that every subset of states having a substantial stationary probability will be reached with high probability during feasible sampling periods $n$. This insures that within the sampling period $n$, the stationary distribution of states is approximated without leaving out important subsets of states. Work is needed to make this requirement more precise in terms of the transition probabilities of the underlying model.

(3) The autocorrelation of the output variable is geometric or of the "decay" type, *i.e.*, if

$$E(X_t^2) = \sigma^2 + \mu^2 \quad \text{and} \quad E(X_t X_{t+1}) = \rho\sigma^2 + \mu^2$$

then

$$E(X_t X_{t+K}) = \rho^k \sigma^2 + \mu^2 \quad . \tag{B-1}$$

This is more flexible than assuming no autocorrelation and it agrees with observed output in several queuing models. The validity of it can be tested by estimating $E(X_t X_{t+k})$ for $k = 2, 3, \ldots$ . This assumption will not hold when periodic effects occur.

(4) At least 100 values of $X_t$ are recommended. Compensation for small sample sizes is not included in the procedures. Normally $n < 100$ anyway.

The output time series $X_t$, $1 \leq t \leq n$, is treated apart from other output time series. It is not necessarily the raw output, but usually is an average of $L$ successive output values $y_{t1}, y_{t2}, \ldots, y_{tL}$:

$$X_t = \frac{1}{L} \sum_{j=1}^{L} y_{tj} \tag{B-2}$$

This reduces computer time for statistical computation which is proportional to $1/L$. Values of $L = 5$ or 10 are recommended. As an upper restriction on $L$, it should not be desirable to stop simulation before 100 groups of $L$ events have occurred.

Some definitions are needed.

$\sigma^2(X)$ and $\sigma^2(\bar{X})$ are the variances of $X_t$ and $\bar{X}$.

$S^2(X)$ and $S^2(\bar{X})$ are sample variances of $X_t$ and $\bar{X}$.

$$\bar{X}_{n1}^{k} = \frac{1}{n-k} \sum_{t=k}^{n} X_t, \qquad \bar{X}_{n2}^{k} = \frac{1}{n-k} \sum_{t=1}^{n-k} X_t \tag{B-3}$$

$$C_k = \sum_{t=1}^{n-k} (X_t - \bar{X}_{n1}^{k})(X_{t+k} - \bar{X}_{n2}^{k}) \tag{B-3a}$$

$$\rho_k^k = \frac{C_k}{(n-k)S^2(X)} \quad \text{is the sample } k^{\text{th}} \text{ order autocorrelation} \quad . \tag{B-3b}$$

$$D_k = \frac{n-2k}{(n-k)^2} \left[ \frac{1+\rho_k}{(1-\rho_k)^2} \right] \quad \text{for} \quad K = 1, 2, 3, \ldots \quad . \tag{B-3c}$$

Then the confidence interval about $\bar{X}$ is $\bar{X} \pm aS(\bar{X})$ where $a$ is determined by the level of confidence level and

$$S^2(\bar{X}) = \frac{S^2(X)}{n^2} \left[ n + 2 \sum_{k=1}^{n-1} (n-k)\rho_E^k \right] < \frac{S^2(X)}{n} \left[ \frac{1+\rho_E}{1-\rho_E} \right], \qquad \rho_E = \rho_1 + (1-\rho_1)D_1. \tag{B-4}$$

For large $n$, this is almost an equality.

Before obtaining a confidence interval, one needs a "reliable" estimate of $\rho$. It is found that

$$\rho_E = \rho_1 + (1 - \rho_1)D_1 \tag{B-5}$$

is almost unbiased, especially when $D_1$ is small. Criteria considered for a good estimate of $\rho$ are

$$\text{a.} \quad \left| E\left(\frac{1 + \rho_E}{1 - \rho_E}\right) - \frac{1 + \rho}{1 - \rho} \right| < h\left(\frac{1 + \rho}{1 - \rho}\right), \qquad h > 0 \tag{B-6a}$$

$$\text{b.} \quad P_r\left[\frac{1 + \rho}{1 - \rho} > h'\left(\frac{1 + \rho_E}{1 - \rho_E}\right)\right] < 0.5, \qquad h' > 1 \tag{B-6b}$$

If $h$ is small, bias is small. If $h'$ is small, variation is small.

When $D_1 < 0.1$ and $n > 100$, it is probably true that $h < 0.1$ and $h' < 1.5$, although no proof of this has been found.

To see why $D_1$ is involved, consider

$$\frac{E(C_k)}{(n - k)\sigma^2(X)} \tag{B-7}$$

and particularly consider $k = 1$.

$$\frac{E(C_k)}{(n - k)\sigma^2(X)} = \rho^k - \frac{n - 2k}{(n - k)^2}\left(\frac{1 + \rho}{1 - \rho}\right) + \left[\text{terms in } \left(\frac{\rho^{k+1}}{(n - k)^2}\right)\right]$$

$$+ \left[\text{terms in } \frac{\rho^{n-3k}}{(n - k)^2}\right], \qquad k = 1, 2, 3, \ldots \tag{B-8}$$

For $n > 100$, and $k$ small, the last two terms are insignificant. The bias is in the second term which can be approximated by $(1 - \rho_1)D_1$. However, since the bias does depend on $\rho$ which we are estimating, a stable solution of this equation occurs only for small values of the second term relative to $1 - \rho$. $D_1$ is a sample estimate of the second term over $1 - \rho_1$. In this way, $D_1$ and $h$ are related.

Further, for $X_t$ normally distributed an approximate 0.95 probability bound on $\rho$ was obtained for cases where $D_1 = 0.1$. It is shown in the next section that $h' < 1.5$ except when $n < 100$. This partially substantiates the relationship between $D_1$ and $h'$.

B. Variance of the Correlation

The stopping rule depends upon finding a relatively unbiased estimate of $\rho$, the correlation. Whether this provides an estimate of $[(1 + \rho)/(1 - \rho)]$ with small variance is the critical question. This section shows that for $D = 0.1$, and $\rho > 1/2$, assuming normally distributed output,

$$P_r \left[ \left( \frac{1 + \rho}{1 - \rho} \right) > h \left( \frac{1 + \rho_E}{1 - \rho_E} \right) \right] < 0.05 \quad \text{with } h \approx 1.5 \quad . \tag{B-9}$$

This would seem to provide a reliable estimate of $[(1 + \rho)/(1 - \rho)]$. Note that $h$ is a function of the stopping rule and $D$ in particular. Other values of $D$ could be used, but only $D = 0.1$ is investigated here.

First, it can be reasonably assumed that

$$X_i = \frac{1}{L} \sum_{j=1}^{L} Y_{ij}$$

is normally distributed, for large values of $L$. For $L = 5$ or 10, this is frequently not a very good assumption, but this seems to be the only case for which the distribution of correlation has been analyzed.

Then the key to this development is a result of Fisher that for normal variates $X$ and $Y$, $Z = 1/2 \log_\rho [(1 + \rho_{XY})/(1 - \rho_{XY})]$ is normally distributed with variance $\sigma_Z^2 = 1/(n - 3)$, where $n$ is the sample size. In this case each pair $(X_t, Y_t)$, $t = 1, 2, \ldots n$ is independent of the others. In serial correlation, $Y_t = X_{t+1}$. The estimate of serial correlation is based on correlated observations, but with approximately the same distribution as $Z$ with $\sigma_Z^2 = 1/(n - 3)[1 + \rho_z)/(1 - \rho_z)]$.

An estimate of $\rho_z$, the correlation between $X_t X_{t-1}$ and $X_{t+1} X_t$ is needed. It turns out that $\rho_z < \rho_E$. Derivation:

Under the geometric assumption used to derive $\rho_E$, $K_t$ is given by

$$X_t = \rho X_{t-1} + \epsilon_t$$

where

$$\epsilon_t \sim n\left(0, \frac{\sigma^2}{1 - \rho^2}\right), \qquad \text{all } t$$

$$X_t \sim n(0, \sigma^2) \qquad \text{all } t \quad ,$$

and

$$E\epsilon_{t+K}\epsilon_t = 0, \qquad K \neq 0$$

$$E\epsilon_{t+K}X_t = 0 \qquad K \neq 0 \quad , \tag{B-10}$$

assuming mean 0 is easiest but completely general. In general, assume that $E(X_t^4) = K\sigma^4$. ($K = 3$ for the normal distribution.) Then

$$E[X_tX_{t-1}]^2 = E[(\rho X_{t-1} + \epsilon_t)X_{t-1}]^2 = E[\rho^2X_{t-1}^4 + 2\rho X_{t-1}^3\epsilon_t + \epsilon_t^2X_{t-1}^2]$$

$$= K\rho^2\sigma^4 + (1 - \rho^2)\sigma^4$$

$$E[(X_tX_{t-1})(X_{t+1}X_t)] = E[(\rho X_{t-1} + \epsilon_t)^2(\rho^2X_{t-1} + \rho\epsilon_t + \epsilon_{t+1})X_{t-1}]$$

$$= E[(\rho^2X_{t-1}^3 + 2\rho X_{t-1}^2\epsilon_t + X_{t-1}\epsilon_t^2$$

$$\cdot (\rho^2X_{t-1} + \rho\epsilon_t + \epsilon_{t+1})]$$

$$= K\rho^4\sigma^4 + 2\rho^2(1 - \rho^2)\sigma^4 + \rho^2(1 - \rho^2)\sigma^4$$

$$= \rho^2\sigma^4[\rho^2K + 3(1 - \rho^2)]$$

$$C_z = E[(X_tX_{t-1})(X_{t+1}X_t)] - \{E[X_tX_{t-1}]\}^2$$

$$= \rho^2\sigma^4(\rho^2K + 3)(1 - \rho^2) - \rho^2\sigma^4$$

$$\rho_z = \frac{C_z}{\sigma_z^2} = \rho^2\left[\frac{(K - 3)\rho^2 + 2}{(K - 2)\rho^2 + 1}\right] \tag{B-11}$$

When is $\rho_z < \rho$?

$$\frac{(K-3)\rho^2 + 2}{(K-2)\rho^2 + 1} < \frac{1}{\rho}$$

$$(K-3)\rho^3 - (K-2)\rho^2 + 2\rho - 1 < 0$$

$$(\rho - 1)(K-3)\rho^2 + 1 - \rho < 0, \qquad \text{so } K \qquad . \tag{B-12}$$

For the normal distribution, $K = 3$ satisfies this inequality, so that estimating $\rho_z$ by $\rho_E$ will be conservative. Note that for $\rho > 4$, $\rho_z$ is close to $\rho$, and for small $\rho$, we have $\rho_z \approx \rho^2$. (For other common distributions $K > 3$, *i.e.*, for exponential $K = 9$.)

Now given an estimate $\rho_E$, the variance of $z$ is estimated by $(1 + \rho_E)/[(n-3)(1-\rho_E)]$ and the upper critical point of $z$ for $\alpha = 0.05$ is

$$z + \frac{2}{\sqrt{n-3}} \left(\frac{1+\rho_E}{1-\rho_E}\right)^{\frac{1}{2}} \tag{B-13}$$

This can be inverted to yield the upper critical value of $\rho$ called $\bar{\rho}$.

$$\bar{\rho} = \frac{\left(\frac{1+\rho_E}{1-\rho_E}\right) e^{\sqrt{\frac{4}{n-3}}\frac{1+\rho_E}{1-\rho_E}} - 1}{\left(\frac{1+\rho_E}{1-\rho_E}\right) e^{\sqrt{\frac{4}{n-3}}\frac{1+\rho_E}{1-\rho_E}} + 1} \qquad . \tag{B-14}$$

For $\rho_E = 0.2, 0.4, 0.6, 0.8, 0.9, 0.95$, these values are given in Table B-I, from which $h = 1.5$ is derived empirically.

For small $\rho$, a good estimate is guaranteed by requiring $n > 100$. For $\rho = 0$, as an example,

$$P_r\{\rho_E > 0.2\} < 0.05 \quad \text{if} \quad n = 100, \tag{B-15}$$

because this yields 100 independent observations on $\rho$, and this distribution has been worked out for normal variates.[14]

Table B-1

APPROXIMATE UPPER CRITICAL VALUES OF $[(1 + \rho)/(1 - \rho)]$ FOR $\alpha = 0.05$, $P\{[(1 + \rho)/(1 - \rho)] > h[(1 + \rho_E)/(1 - \rho_E)]\} = 0.05$, Assuming Normal Variates.

| A | B | | | | C | | | |
|---|---|---|---|---|---|---|---|---|
| $\rho_E$ | $\frac{1 + \rho_E}{1 - \rho_E}$ | $n = \frac{(B)10}{1 - \rho_E}$ | $\sigma_z^2 = \frac{B}{n - 3}$ | $2\sigma_z$ | $e^{2\sigma_z}$ | $\frac{BC - 1}{BC + 1} = \overline{\rho}$ | $\frac{1 + \overline{\rho}}{1 - \overline{\rho}}$ | $h^*$ |
| 0.2 | 1.5 | 18.7 | 0.1 | 0.64 | 1.9 | $\frac{1.75}{3.75} = 0.47$ | 3.0 | 2.0 |
| 0.4 | 2.33 | 37 | 0.06 | 0.50 | 1.65 | $\frac{2.84}{4.84} = 0.59$ | 4.0 | 1.7 |
| 0.6 | 4 | 100 | 0.04 | 0.4 | 1.5 | $\frac{5}{7} = 0.71$ | 5.7 | 1.4 |
| 0.8 | 9 | 450 | 0.02 | 0.28 | 1.33 | $\frac{11}{13} = 0.85$ | 12 | 1.5 |
| 0.9 | 19 | 1900 | 0.01 | 0.2 | 1.22 | $\frac{22}{23} = 0.92$ | 24 | 1.3 |
| 0.95 | 39 | 7800 | 0.005 | 0.14 | 1.15 | $\frac{44}{45} = 0.956$ | 45 | 1.15 |

Note that no upward adjustment for bias in $\rho_E$ is made. Also, no downward adjustment is made for the known difference between $\rho_E$ and $\rho_z$. For $\rho > 1/2$, these adjustments are cancelling and negligible.

* For small $\rho$, $\rho_z \approx \rho_E^2$ and any adjustment should be downward, so that $h$ is overestimated here.

Note that for $\rho_E < 0.2$, $\frac{1 + \rho_E}{1 - \rho_E} < 1.5$ as desired.

Further, even with $\rho = 0$, one would expect correlation between $X_t X_{t+1}$ and $X_{t+1} X_{t+2}$. But these are independent when $\rho = 0$, so that 100 independent observations are obtained for $n = 100$, $\rho = 0$.

### C. Variance of $\rho_E = \left(\rho_K^K\right)^{1/K}$

For $K > 1$, the estimate $\rho_K^K$ of $\rho^K$ has the same form and variance as $\rho_1$ does as an estimator of $\rho$. But when the $K$th root is found, the variance is magnified as $K$ increases; $\left(\rho_K^K\right)^{1/K}$ becomes a very poor estimator of $\rho$. Experience has shown that this can be a serious effect. Thus, use of $\left(\rho_K^K\right)^{1/K}$ for $> 1$ is not recommended for small sample sizes, despite the risk incurred if the geometric correlation assumption is invalid.

### D. Experience with Confidence Interval Procedure

It was once felt that $\rho_1, \rho_2, \ldots, \rho_5$ might be good estimators of $\rho$. However, we see that only $\rho_1$ is good for small samples. This is because $\rho_1, \rho_2^2, \rho_3^3, \ldots, \rho_5^5$ have the same variation. For example, $\left(\rho_5^5\right)^{1/5}$ has more variation than $\rho_1$. But for large sample sizes, the validity of the geometric assumption can be tested using the sample estimates of higher correlations.

When only $\rho_1$ was estimated and $L = 5$, all of the statistical computation added about 20% to the basic running time of a simulation. If $L = 10$, the computation added would be about 10% instead of 20%, so an appreciable gain is possible when large grouping is feasible.

This procedure has thus far agreed with theoretical results when applied to both input and output variables. In several instances the designed mean of exponential input was verified within a 0.034% interval about the observed average. Similarly, within 10% of the observed average of output variables, theoretical values were found as specified from sample confidence intervals. In some cases the theoretical value is close to the boundary of the confidence interval so that the intervals do not seem to be grossly in error.

*APPENDIX C*

## CHOPPING RULES AND BIAS FROM INITIAL STATES

If the initial state is not selected at random from the true distribution of states when the system is in steady-state the choice will introduce some bias into the resultant average statistics. It is desirable to determine and possibly eliminate this bias.

Past suggestions have been:

(1) Ignore this effect and start from zero, no elements in system.

(2) Use a preliminary run and choose the initial state at random from the end section of the preliminary run.

(3) Drop the first part of the run from average statistics.

We are using a stopping rule that detects high correlation, which would accompany a large effect from the initial state. Thus, the stopping rule tends to force a large sample to minimize the effect of the initialization. Experience also indicates that we can ignore chopping rules.

The steady state probability of "idle state" occurring is easily determined as $\rho = (\mu - \lambda)/\mu$ where $\lambda$ is the arrival rate (all priorities) and $\mu$ is the service rate (all priorities). This holds true except for the case of preemptive priorities and state-dependent arrival and service rates. In these cases, conservative estimates can be found.

Then for samples of 100, and $\rho < 0.9$, more than 10 occurrences of the initial state would be expected in the run. Little bias would be expected when starting from such a common state. So there is justification for ignoring the effect of initialization. The use of a preliminary run to generate random states produces several problems:

(1) How long does the preliminary run have to be? This is essentially the same as asking about the size of the initialization effect. Just knowing that a preliminary run will decrease the effect is no more information than knowing that the system will approach a steady-state condition.

(2) If several initial states are to be recorded from the preliminary run, for future use, a complete state description of the system must be recorded. On the other hand, in truncating the first part of a run, the state of the system is naturally generated in transition.

The main disadvantage in chopping off the first part of a run is that the chopping point must depend only on statistics preceding the chopping point or else retroactive adjustments are needed.

We propose here a method of retroactive chopping which requires minor retroactive adjustments and which adaptively determines the chopping point.

The notation of the stopping rule will be used.

Assume that the stopping rule criteria have been met for all statistical averages, so that the run is about to terminate on that basis. At this point, a check for initialization bias is made on each statistical average. After each sequence is chopped, the stopping rule is applied again. If more observations are needed, the same cycle will be repeated until the stopping rule is satisfied immediately after chopping.

Now consider the chopping procedure for one sequence $X_1, X_2, \ldots, X_n$. Here $n$ is the number of observations when the stopping rule is satisfied, and the statistics $\bar{X}_n$, $S_n^2$, $S_{\bar{X}_n}^2$ and $\rho_E$ are determined.

Consider an intermediate point $m$ where statistics $\bar{X}_m$ and $S_m^2$ are stored.

The sequence $X_{m+1}, X_{m+2}, \ldots, X_n$ has

$$\bar{X}_{n-m} = \frac{n\bar{X}_n - m\bar{X}_m}{n - m} ,$$

$$S_{n-m}^2 = \left[(n-1)S_n^2 - (m-1)S_m^2 - (n-m)\bar{X}_{n-m}^2 + n\bar{X}_n^2 - m\bar{X}_m^2\right] \frac{1}{n-m-1} . \tag{C-1}$$

It is unnecessary to test whether the mean $\mu_m < \mu_{n-m}$ significantly if $\bar{X}_m > \bar{X}_{n-m}$. But if $\bar{X}_m < \bar{X}_{n-m}$, assuming equal variances for both sequences, a test statistic $t$ normally distributed can detect significant effects of initialization. If starting from the "idle state" would inflate values of $X$, then the test would be made in the other direction.

Let

$$t = \frac{\bar{X}_{n-m} - \bar{X}_m}{\left(\frac{1}{n-m} + \frac{1}{m}\right)^{\frac{1}{2}} \left(\frac{(n-1)S_n^2 - (n-m)\bar{X}_{n-m}^2 + n\bar{X}_n^2 - m\bar{X}_m^2}{n-2}\right)^{\frac{1}{2}} \left(\frac{1+\rho_E}{1-\rho_E}\right)^{\frac{1}{2}}} \tag{C-2}$$

If $\bar{X}_{n-m}$ and $\bar{X}_m$ are normally distributed with equal variances, $t$ has the $t$-distribution with $n-2$ degrees of freedom. But for $n > 100$, this is very close to the normal distribution. So for $t > 2$, we can reject the hypothesis that there is no initialization effect and be wrong only 2.5% of the time. Because of correlation, this will be a conservative test, that is, $t$ will tend to be smaller as correlation increases.

If $t > 2$, then the sequence can be chopped at $m$, leaving $X_{m+1}$, $X_{m+2}$, ..., $X_n$ with statistics

$$\bar{X}_{n-m},\ S_{n-m}^2,\ \rho_E$$

and

$$S_{\bar{X}_{n-m}}^2 = \frac{S_{n-m}^2}{n-m}\left[\frac{1+\rho_E}{1-\rho_E}\right] . \tag{C-3}$$

Note that $\rho_E$, as determined for the full sequence, has not been changed. If $|\bar{X}_{n-m} - \bar{X}_m| < 1/10(S_{\bar{X}_n})$, the chopping may not be worth the trouble.

This test should not be made often, because the probability of chopping wrongly tends to accumulate as a multiple of the number of tests ($T$), *i.e.*, $P(t > 2$ in one of $T$ tests$|$no initial effect$) \approx 0.025\ T$. It is recommended that if $n'$ is the number of samples between tests of the stopping rule, the points $m$ could be

$$\frac{n'}{10},\ \frac{n'}{2},\ n',\ 2n',\ 4n',\ 8n',\ \ldots,$$

such that at these points, $\bar{X}_m$ and $S_m^2$ are stored.

The chopping decision at point $m$ is made only if $n \geq 2m$. The greatest value of $m$ with $t > 2$ is the chopping point, and the procedure continues with the remaining $n - m$ points. This means not only revising the $n$-point statistics, but also scrapping all stored values at $m$-points less than $n$, and picking up new $m$ points as they occur in the new $(n - m)$ sequence. That is, if $n' = 100$, $n = 500$ and the sequence is chopped at 50, $n - m = 450$ and all $m$ points are discarded. The first one picked up will be at $n - m = 800$, and cannot be used unless $n - m$ exceeds 1600.

## REFERENCES


1. "473-L Simulation," Monthly Status Report, Contract AF 19(628)-3843, Information System Operations, General Electric Company, Washington, D.C. (February 1964).

2. Nee, D. "Application of Queuing Theory to Information System Design," Final Report, Contract AF 19(628)-2901, Stanford Research Institute, Menlo Park, California (June 1964).

3. Cox, R. E., "Traffic Flow in an Exponential Delay System with Priority Categories," Proc. IEEE, (London) 102, Pt. B, (1955), p. 815-818.

4. Conway, "Some Tactical Problems in Digital Simulation," Management Science, Vo. 10, No. 1, p. 47.

5. Hammersley and Handcomb, Monte Carlo Methods, Methuen Monographs, and Wiley, 1960.

6. Clark, C. E., The Utility of Statistics of Random Numbers, Ops. Res. Vol. B, 1060, pp. 185-195.

7. Page, E. E., "Simulation in Queuing Systems," Ops. Res. Vol. 13, No. 2, pp. 300-305.

8. Kahn and Marshall, Methods of Reducing Sample Size in Monte Carlo Computations, Ops. Res. Vol. 1, 1953, pp 263-278.

9. Ehrenfeld and Ben-Tuvia, "The Efficiency of Statistical Simulation Procedures," Technometrics, Vol. 4, 1962, pp. 257-275.

10. Tocher, K. D., The Art of Simulation, D. Van Nostrand Co., Inc., Princeton, New Jersey, 1963.

11. Bellman, An Introduction to Matrix Analysis.

12. Burke, Paul J., "The Output of Queuing Systems," pp. 699-704 (1956).

13. Harrison White and Lee G. Christie, "Queuing with Preemptive Priority or with Breakdown," Opers. Res. 6, 79-96 (1958).

14. Ostel, Statistics in Research, p. 459.

UNCLASSIFIED

Security Classification

## DOCUMENT CONTROL DATA - R&D

*(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)*

1 ORIGINATING ACTIVITY *(Corporate author)*

Stanford Research Institute
Menlo Park, California

2a REPORT SECURITY CLASSIFICATION

Unclassified

2b GROUP

N/A

3 REPORT TITLE

APPLICATION OF QUEUING THEORY TO INFORMATION SYSTEM DESIGN (Final Report)

4 DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

None

5 AUTHOR(S) *(Last name, first name, initial)*

Nee, David

| 6 REPORT DATE | 7a TOTAL NO. OF PAGES | 7b NO. OF REFS |
|---|---|---|
| November 1965 | 102 | 14 |

| 8a CONTRACT OR GRANT NO. | 9a ORIGINATOR'S REPORT NUMBER(S) |
|---|---|
| AF 19(628)-4341 | ESD-TR-65-577 |
| b. PROJECT NO. SRI 5187 | |
| c. | 9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)* |
| d. | |

10 AVAILABILITY/LIMITATION NOTICES

This document is subject to special export controls and each transmittal to foreign governments or foreign nationals may be made only with prior approval of Hq ESD (ESTI).

| 11 SUPPLEMENTARY NOTES | 12 SPONSORING MILITARY ACTIVITY |
|---|---|
| None | Directorate of Computers, Electronic Systems Division, AFSC, USAF, L. G. Hanscom Field, Bedford, Mass. 01731 |

13 ABSTRACT

This research project was undertaken to evaluate the applicability of queuing theory—in particular the priority queuing theory—to the evaluation of priority queuing situations in military information systems.

The queuing theory has been applied to the evaluation of the queuing situations in the 473-L System. The waiting-time distribution for a single-server, head-of-the-line, priority queuing model has been evaluated. The application of "Variance Reduction Method" to the Simulation of Priority Queuing Systems has been investigated. Finally, guides and procedures for the application of queuing models in structuring information systems were outlined.

DD FORM 1 JAN 64 1473

UNCLASSIFIED

Security Classification

| 14 KEY WORDS | LINK A | | LINK B | | LINK C | |
|---|---|---|---|---|---|---|
| | ROLE | WT | ROLE | WT | ROLE | WT |
| Information Systems<br>Models<br>Simulation<br>Design<br>Statistical Analysis<br>Queuing Theory<br>Operations Research<br>Probability | | | | | | |

## INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2*a*. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2*b*. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7*a*. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7*b*. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8*a*. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8*b*, 8*c*, & 8*d*. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9*a*. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9*b*. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, *or (U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, roles, and weights is optional.